Landscape Analysis

Big Data Tools

Authors	Andrea Manzi, Diego Scardaci
Technology	Big Data Tools
Last update	09/02/2021
Status	Final

DOCUMENT LOG

Issue	*Date*	*Comment*	*Author*
v0.1	08/10/2020	First version	Andrea Manzi
v0.2	22/10/2020	Revised section 2 and 3	Andrea Manzi
v0.3	20/11/2020	Added Spark and NoSQL key-value on section 3, worked on the other sections,	Andrea Manzi
v0.4	23/11/2020	Finalized first full draft	Andrea Manzi
v0.5	23/12/2020	Reviewed version	Diego Scardaci
v0.6	08/01/2021	version addressing review comments	Andrea Manzi
V0.7	09/02/2021	New reviewed version	Diego Scardaci
V1.0	09/02/2021	Final version addressing review comments	Andrea Manzi

TERMINOLOGY

For the purpose of this document, the following terms and definitions apply:

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", “MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119. For a complete list of term definitions see the EGI Glossary (http://wiki.egi.eu/wiki/Glossary).

Contents

Executive Summary

The document reports a landscape analysis of the Big Data tools, starting from a description of the general use cases and fields of applicability (chapter 2), an overview of the available tools classified in 4 areas according to the phase of the big data management lifecycle (chapter 3) and an analysis on the current adoption and usage of the technologies (chapter 4). After these chapters the document focuses on standardisation activities and policies (chapter 5), projects, initiatives, partnerships and technology providers (chapters 6 and 7) and possible integration scenarios in the EGI infrastructure (chapter 8).

Big Data refers to the non-traditional strategies and technologies needed to collect, organize, process, and gather insights from large datasets. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing has greatly expanded in recent years.

In addition, the hype of Big Data services and tools is now linked to other technological trends such as the Internet of Things (IoT) and Artificial Intelligence(AI)/ Machine Learning (ML), and it's focusing more on what is referred to analytics of data-in-motion rather than data-at-rest.

Development of all big data related technologies and approaches has mainly been driven by non-EU companies, and in particular lots of big data tools and services are provided by the big US public cloud providers. Recently though in Europe, the EC with the publication of the European Data Strategy[1] is planning to become a leader in a data-driven society hence also providing big data tools services in the EU. A number of related EU projects will be funded to implement this strategy and EGI could play an important role bringing its experience on dealing with very large datasets in extreme distributed environments.

[1] https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en

1 Introduction

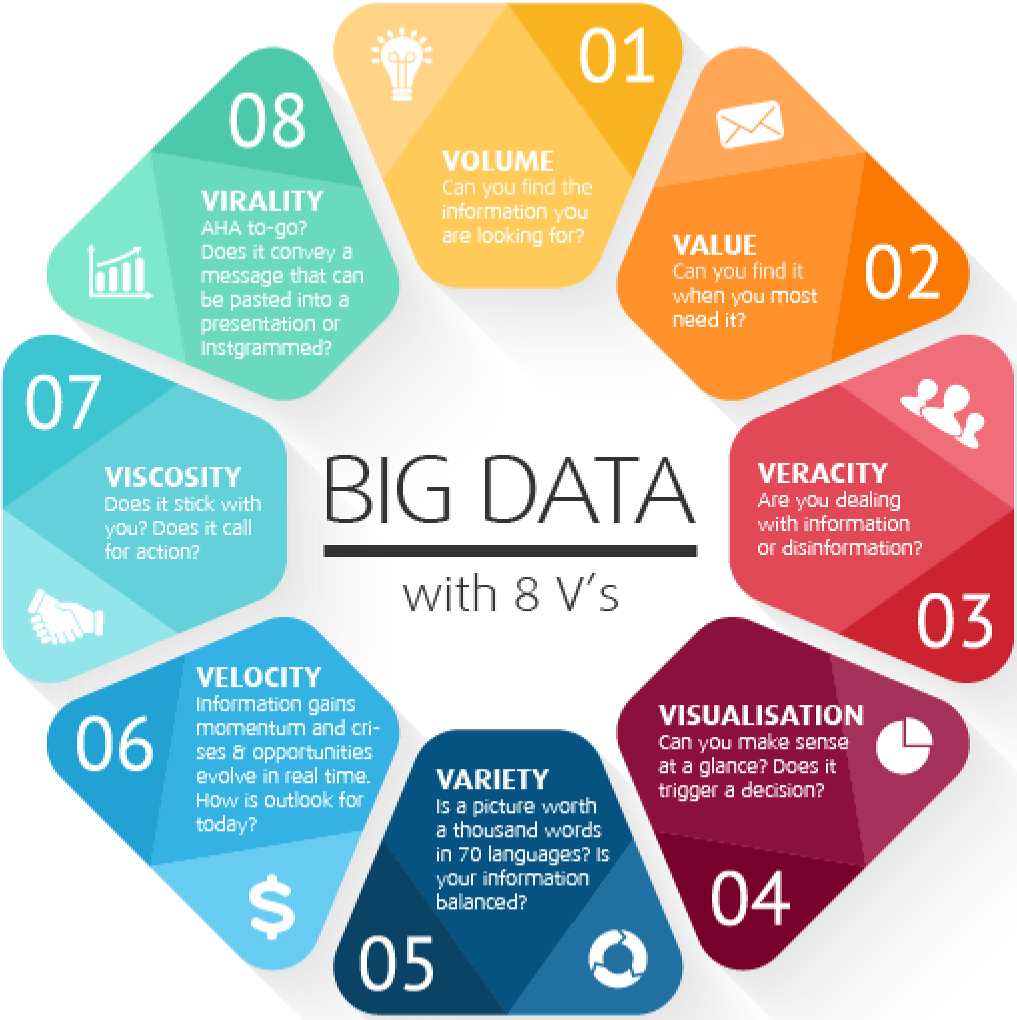
Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. Therefore, big data often includes data with sizes that exceed the capacity of traditional software to process within an acceptable time and value.

Current usage of the term big data tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet searches, fintech, urban informatics, and business informatics. Scientists encounter limitations in e-Science work in a wide set of disciplines, including meteorology, genomics, connectomics, complex physics simulations, biology and environmental research.

Data sets grow rapidly, to a certain extent because they are increasingly gathered by cheap and numerous information-sensing Internet of things devices such as mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exabytes (2.5×2^{60} bytes) of data are generated. Based on an IDC report prediction by 2025 there will be 163 zettabytes of data.

What qualifies as being "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration".[2]

[2] https://en.wikipedia.org/wiki/Big_data


01
VOLUME
Can you find the information you are looking for?
02
VALUE
Can you find it when you most need it?
03
VERACITY
Are you dealing with information or disinformation?
04
VISUALISATION
Can you make sense at a glance? Does it trigger a decision?
05
VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?
06
VELOCITY
Information gains momentum and cri-ses & opportunities evolve in real time. How is outlook for today?
07
VISCOSITY
Does it stick with you? Does it call for action?
08
VIRALITY
AHA to-go? Does it convey a message that can be pasted into a presentation or Instgrammed?
BIG DATA
with 8 V's

The above image depicts the 8 V's which best define Big Data.[3] Among those the peculiar one that are described everywhere are:

- **Volume** defines the huge amount of data that is produced each day by companies, for example. The generation of data is so large and complex that it can no longer be saved or analyzed using conventional data processing methods.
- **Variety** refers to the diversity of data types and data sources. 80 percent of the data in the world today is unstructured and at first glance does not show any indication of relationships. Thanks to Big Data such algorithms, data is able to be sorted in a structured manner and

[3] https://www.zarantech.com/blog/effective-way-handle-big-data/

examined for relationships. Data does not always comprise only conventional datasets, but also images, videos and speech recordings.

- **Velocity** refers to the speed with which the data is generated, analyzed and reprocessed. Today this is mostly possible within a fraction of a second, known as real time.

2 Demand, use cases and fields of applicability

2.1 Demand

2.1.1 Education Industry[4]

Education industry is flooded with huge amounts of data related to students, faculty, courses, and results. Proper study and analysis of this data can provide insights which can be used to improve the operational effectiveness and working of educational institutes. Following a list of some of the fields in the education industry that have been transformed by big data-motivated changes:

- Customized and Dynamic Learning Programs
- Reframing Course Material
- Grading Systems
- Career Prediction

2.1.2 Healthcare

Healthcare is another industry which is bound to generate a huge amount of data. Following are some of the ways in which big data has contributed to healthcare:

- Big data reduces costs of treatment since there are less chances of having to perform unnecessary diagnosis.
- It helps in predicting outbreaks of epidemics and also in deciding what preventive measures could be taken to minimize the effects of the same.
- It helps avoid preventable diseases by detecting them in early stages. It prevents them from getting any worse which in turn makes their treatment easy and effective.
- Patients can be provided with evidence-based medicine which is identified and prescribed after doing research on past medical results.

2.1.3 Government Sector

Governments come face to face with a very huge amount of data on an almost daily basis. The reason for this is, they have to keep track of various records and databases regarding their citizens,

[4] https://intellipaat.com/blog/7-big-data-examples-application-of-big-data-in-real-life/

their growth, energy resources, geographical surveys, and many more. All this data contributes to big data. The proper study and analysis of this data, hence, helps governments in endless ways. Few of them are as follows:

Welfare Schemes

- In making faster and informed decisions regarding various political programs
- To identify areas that are in immediate need of attention
- To stay up to date in the field of agriculture by keeping track of all existing land and livestock.
- To overcome national challenges such as unemployment, terrorism, energy resources exploration, and much more.

Cyber Security

- Big Data is hugely used for deceit recognition.
- It is also used in catching tax evaders.

2.1.4 Media and Entertainment Industry

With people having access to various digital gadgets, generation of large amounts of data is inevitable and this is the main cause of the rise in big data in the media and entertainment industry.

Other than this, social media platforms are another way in which a huge amount of data is being generated. Although, businesses in the media and entertainment industry have realized the importance of this data, and they have been able to benefit from it for their growth.

Some of the benefits extracted from big data in the media and entertainment industry are given below:

- Predicting the interests of audiences
- Optimized or on-demand scheduling of media streams in digital media distribution platforms
- Getting insights from customer reviews
- Effective targeting of the advertisements

2.1.5 Weather Patterns

There are weather sensors and satellites deployed all around the globe. A huge amount of data is collected from them, and then this data is used to monitor the weather and environmental conditions.

All of the data collected from these sensors and satellites contribute to big data and can be used in different ways such as:

- In weather forecasting

- To study global warming
- In understanding the patterns of natural disasters
- To make necessary preparations in the case of crises
- To predict the availability of usable water around the world

2.1.6 Transportation

Since the rise of big data, it has been used in various ways to make transportation more efficient and easy. Following are some of the areas where big data contributes to transportation.

- Route planning: Big data can be used to understand and estimate users' needs on different routes and on multiple modes of transportation and then utilize route planning to reduce their wait time.
- Congestion management and traffic control: Using big data, real-time estimation of congestion and traffic patterns is now possible. For example, people are using Google Maps to locate the least traffic-prone routes.
- Safety level of traffic: Using the real-time processing of big data and predictive analysis to identify accident-prone areas can help reduce accidents and increase the safety level of traffic.

2.1.7 Banking

The amount of data in the banking sector is also skyrocketing. Proper study and analysis of this data can help detect any and all illegal activities that are being carried out such as:

- Misuse of credit/debit cards
- Venture credit hazard treatment
- Business clarity
- Customer statistics alteration
- Money laundering
- Risk mitigation

2.1.8 Science

Examples of Big Data in science are the LHC experiments, SKA, NASA, DNA Databases, etc etc, where the amount of data produced by the experiments or simulations is vast and need to be analyzed in distributed infrastructure using Big Data tools and techniques.

2.2 Use cases

2.2.1 Big Data analytics

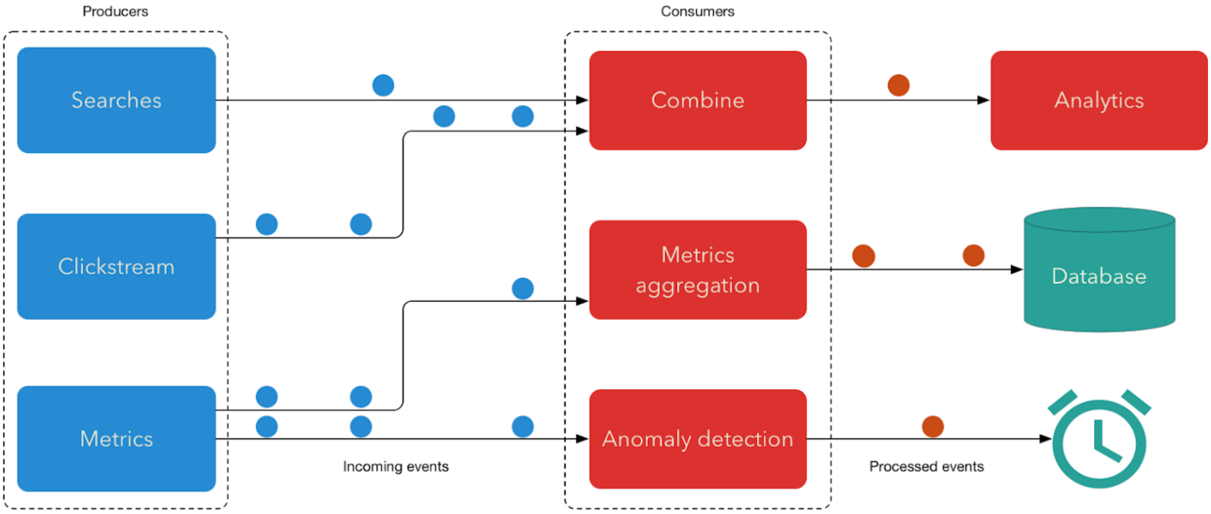
Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes. Analysis of big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable. Businesses can use advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics and natural language processing to gain new insights from previously untapped data sources independently or together with existing enterprise data.

Existing EGI Communities/Projects

- The project EGI-ACE coordinated by EGI, includes some communities with requirements on Big Data Analytics, such as ENES in the domain of climate data analysis.
- The SoBigData RI for Big Data and Social Mining is one of the community exploiting advanced Big Data Analytics tools with links to EGI, which is member of to the SobigData++ project
- RIs member of the ENVRI-FAIR cluster project
- EUHubs4Data, an EC funded project federating European Digital Innovation Hubs (DIHs) with relevant expertise on BigData analytics (EGI is a project member)

2.2.2 Stream processing

Stream processing is the practice of taking action on a series of data at the time the data is created. Stream processing often entails multiple tasks on the incoming series of data (the “data stream”), which can be performed serially, in parallel, or both. This workflow is referred to as a stream processing pipeline, which includes the generation of the data, the processing of the data, and the delivery of the data to a final location. Actions that stream processing takes on data include aggregations (e.g., calculations such as sum, mean, standard deviation), analytics (e.g., predicting a future event based on patterns in the data), transformations (e.g., changing a number into a date format), enrichment (e.g., combining the data point with other data sources to create more context and meaning), and ingestion (e.g., inserting the data into a database).



Existing EGI Communities/Projects

- EUHubs4Data is mentioned again, as some of the DIHs involved are also expert in delivering Stream processing solutions, in particular Real time stream processing linked to IoT and Edge.

2.2.3 Elastic provisioning and scale of Big Data services/tools

Auto-scaling refers to the dynamic addition of nodes to or removal of the nodes from the existing cluster to use the resources effectively. The Quality of Service(QoS) has to be maintained while auto-scaling the resources on-demand. There is a requirement for auto-scaling, for instance Hadoop clusters, when the load increases to adhere to the Service Level Agreements (SLAs).

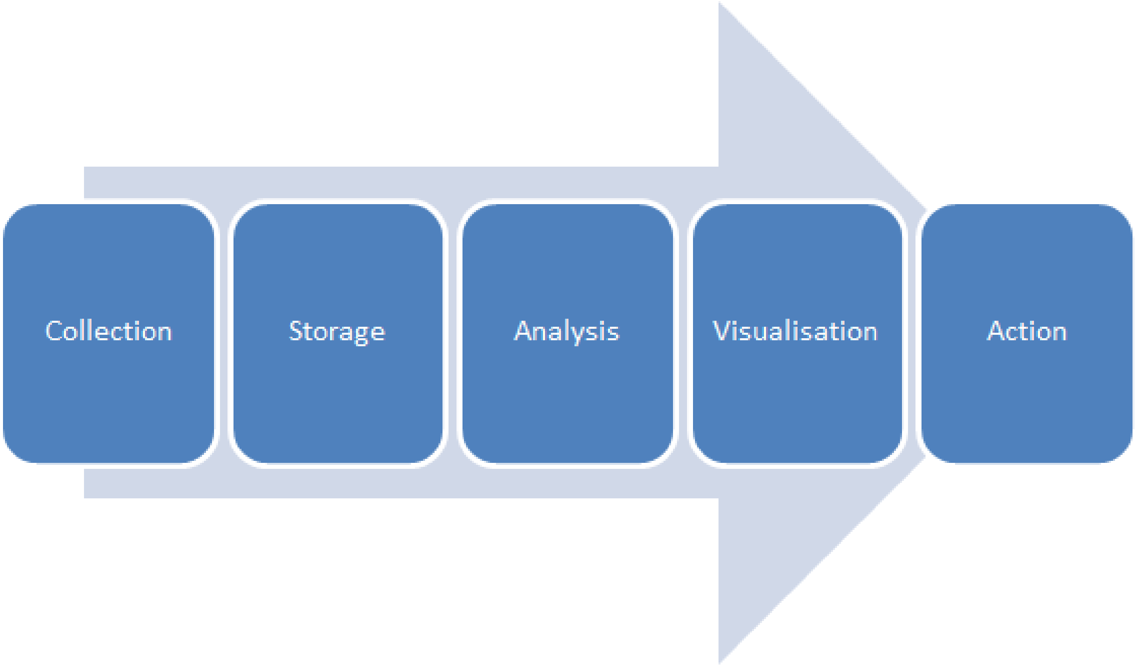
Existing EGI Communities/Projects

- The BD4NRG project is planning to provide big data elastic pipeline orchestrators to be validated through the delivery of predictive and prescriptive edge AI-based big data analytics on 13 large scale pilots, deployed by different energy stakeholders.
- The ENES community has already integrated the automatic deployment and scale of their Big Data analytics tools (ENES Climate Analytics Service (ECAS)) in the EGI infrastructure.

3 Available tools/services

When talking about Big Data tools we should categorize them according to the different phases of the Big Data management lifecycle: Data Collection, Data Storage, Data Analytics and Visualization.

This section shortly describes tools able to support each of these phases.



3.1 Data Collection

Data collection or ingestion, is the process of collecting, filtering and removing any noise from data before they can be stored in any data warehouse or storage system. It adopts adaptive and time efficient algorithms for processing of high value data. Most common solutions make uses of distributed queuing management systems in addition to the real-time data streaming techniques described in section 3.3.3. Some others are making use of technology to import data from databases or streaming files.

Technology	License	Description	Origin
Apache Kafka	Apache License 2.0	Kafka combines three key capabilities: to publish (write) and subscribe to (read) streams of events, including continuous import/export of data from other systems. To store streams of events durably	Global

		and reliably. To process streams of events as they occur or retrospective.	
RabbitMQ	Mozilla Public License	The RabbitMQ server program is built on the Open Telecom Platform framework for clustering and failover. It supports AMQP, STOMP and MQTT protocols.	US
Apache Active MQ	Apache License 2.0	The most popular open source, multi-protocol, Java-based messaging server. The next generation, Artemis, supports AMQP, STOMP and MQTT protocols.	US
Amazon Kinesis Firehose	Proprietary	Ingest real-time data such as video, audio, application logs, website clickstreams, and IoT telemetry data for machine learning, analytics, and other applications.	US
Microsoft Event Hub	Proprietary	Event Hub is a fully managed, real-time data ingestion service. Stream millions of events per second from any source to build dynamic data pipelines and respond to business challenges.	US
Google Pub/Sub	Proprietary	Pub/Sub implement messaging and ingestion for event-driven systems and streaming analytics.	US
Apache Sqoop	Apache License 2.0	Tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases.	Global
Apache flume	Apache License 2.0	Distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It is quite robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms.	Global

3.2 Data Storage

We divided the Data Storage solutions for BigData in 6 categories according to the architecture, implementations and use cases supported.

3.2.1 File systems for Big Data

Technology	License	Description	Origin
HDFS	Apache License 2.0	The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data.	Global
CephFS	LGPLv2.1	The Ceph File System, or CephFS, is a POSIX-compliant file system built on top of Ceph's distributed object store, RADOS. CephFS endeavors to provide a state-of-the-art, multi-use, highly available, and performant file store for a variety of applications, including traditional use-cases like shared home directories, HPC scratch space, and distributed workflow shared storage.	Global
IBM SpectrumScale (GPFS)	Proprietary	GPFS is a clustered file system defined over multiple nodes. It allows users shared access to files that may span multiple disk drives on multiple nodes. UNIX file system utilities are also supported by GPFS. It allows both parallel and serial applications running on different nodes to share data spanning multiple disk drives attached to multiple nodes.	US
GlusterFS	GPLV2 and GPLV3.	Gluster is a scalable network filesystem. Using common off-the-shelf hardware, users can create large, distributed storage solutions for media streaming, data analysis, and other data- and bandwidth-intensive tasks..	Global
Lustre	GPLV2	Lustre is a massively, global, parallel distributed file system, generally used for large scale cluster computing. It provides a high performance file system for computer clusters ranging in size from small workgroup clusters to large-scale, multi-site clusters	Global

3.2.2 Relational Databases

In some use cases it makes sense to still use RDBMS for BigData as they are still good on the volume front, but their fundamental nature makes them ill-suited for velocity and variety. Data does not conform to some kind of predefined schema (unstructured) and stored too fast and too heterogeneously cannot be easily classified for RDBMS purposes. Therefore we just report here shortly the main RDBMS systems: SQL Server, PostgreSQL, MariaDB, Oracle. Public cloud providers deliver RDBMS services like: Google Cloud SQL, Amazon RDS and Azure Cloud Database.

3.2.3 NoSQL Key-value store

From an API perspective, key-value stores are the simplest NoSQL data stores to use. The client can either get the value for the key, assign a value for a key or delete a key from the data store. The value is a blob that the data store just stores, without caring or knowing what's inside; it's the responsibility of the application to understand what was stored. Since key-value stores always use primary-key access, they generally have great performance and can be easily scaled. The key-value database uses a hash table to store unique keys and pointers (in some databases it's also called the inverted index) with respect to each data value it stores. There are no column type relations in the database; hence, its implementation is easy.

Technology	License	Description	Origin
Redis	BSD	Redis is an open source, in-memory data structure store, used as a database, cache and message broker	US
Aerospike	GNU Affero General Public License	Distributed, scalable NoSQL database with ACID support	US
Riak KV	Apache License 2.0	Distributed NoSQL database that is highly available, scalable and easy to operate. It automatically distributes data across the cluster to ensure fast performance and fault-tolerance.	US
Amazon DynamoDB	Proprietary	DynamoDB can handle more than 10 trillion requests per day and can support peaks of more than 20 million requests per second	US
Oracle NOSQL database	Proprietary and	NoSQL Database is a sharded (shared-nothing) system which distributes the data uniformly	US

	community edition	across multiple shards in a cluster. Within each shard, storage nodes are replicated to ensure high availability, rapid failover in the event of a node failure and optimal load balancing of queries	

3.2.4 NoSQL Column based

In column-oriented NoSQL databases, data is stored in cells grouped in columns of data rather than as rows of data. Columns are logically grouped into column families. Column families can contain a virtually unlimited number of columns that can be created at runtime or while defining the schema. Read and write is done using columns rather than rows. Column families are groups of similar data that are usually accessed together. As an example, we often access customers' names and profile information at the same time, but not the information on their orders. The main advantages of storing data in columns over relational DBMS are fast search/access and data aggregation. Relational databases store a single row as a continuous disk entry. Different rows are stored in different places on the disk while columnar databases store all the cells corresponding to a column as a continuous disk entry, thus making the search/access faster. Each column family can be compared to a container of rows in an RDBMS table, where the key identifies the row and the row consists of multiple columns. The difference is that various rows do not have to have the same columns, and columns can be added to any row at any time without having to add them to other rows.

Technology	License	Description	Origin
Apache Cassandra	Apache License 2.0	Free and open-source, distributed, wide column store, NoSQL database management system designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure. Cassandra offers robust support for clusters spanning multiple datacenters, with asynchronous masterless replication allowing low latency operations for all clients.	Global
HBase	Apache License 2.0	This project's goal is the hosting of very large tables , i.e. billions of rows X millions of columns, on top of clusters of commodity hardware. Apache HBase provides Bigtable-like capabilities on top of Hadoop and HDFS.	Global

HyperTable	GNU General Public License 3.0	Hypertable is a high performance, open source, massively scalable database modeled after Bigtable, Google's proprietary, massively scalable database. https://hypertable.com/why_hypertable/hypertable_vs_hbase_2/	Global
Google BigTable	Proprietary	A fully managed, scalable NoSQL database service for large analytical and operational workloads.	US

3.2.5 NoSQL Document based

Document store NoSQL databases are similar to key-value databases in that there’s a key and a value. Data is stored as a value and its associated key is the unique identifier for that value. The difference is that, in a document database, the value contains structured or semi-structured data. This structured/semi-structured value is referred to as a document and can be in XML, JSON or BSON format.

Technology	License	Description	Origin
MongoDB	Community Edition: Versions released prior to October 16, 2018 are published under the AGPL. All versions released after October 16, 2018, including patch fixes for prior versions, are published under the Server Side Public License (SSPL) v1.	MongoDB is a distributed database at its core, so high availability, horizontal scaling, and geographic distribution are built in. MongoDB stores data in flexible, JSON-like documents, meaning fields can vary from document to document and data structure can be changed over time	Global
Apache CouchDB	Apache License 2.0	CouchDB uses multiple formats and protocols to store, transfer, and process its data, it uses JSON to store data, JavaScript as its query language using MapReduce, and HTTP for an API.	Global

RethinkDB	Apache License 2.0	First open-source scalable database built for real time applications. It exposes a new database access model, in which the developer can tell the database to continuously push updated query results to applications without polling for changes.	Global

3.2.6 Graph databases

Graph databases are basically built upon the Entity – Attribute – Value model. Entities are also known as nodes, which have properties. It is a very flexible way to describe how data relates to other data. Nodes store data about each entity in the database, relationships describe a relationship between nodes, and a property is simply the node on the opposite end of the relationship. Whereas a traditional database stores a description of each possible relationship in foreign key fields or junction tables, graph databases allow for virtually any relationship to be defined on-the-fly.

Technology	License	Description	Origin
Neo4j	community edition: GPLv3 and AGPLv3	ACID-compliant transactional database with native graph storage and processing	US
Amazon Neptune	Proprietary	Fully managed graph database service. Amazon Neptune supports graph models Property Graph and W3C's RDF, and their respective query languages Apache TinkerPop Gremlin and SPARQL.	US
OrientDB	Community Edition: Apache License 2	Multi-Model Open Source NoSQL DBMS that combines the power of graphs and the flexibility of documents into one high-performance operational database.	Global
ArangoDB	Apache License 2	Native multi-model database with flexible data models for documents, graphs, and key-values. Provides a convenient SQL-like query language	US-Germany

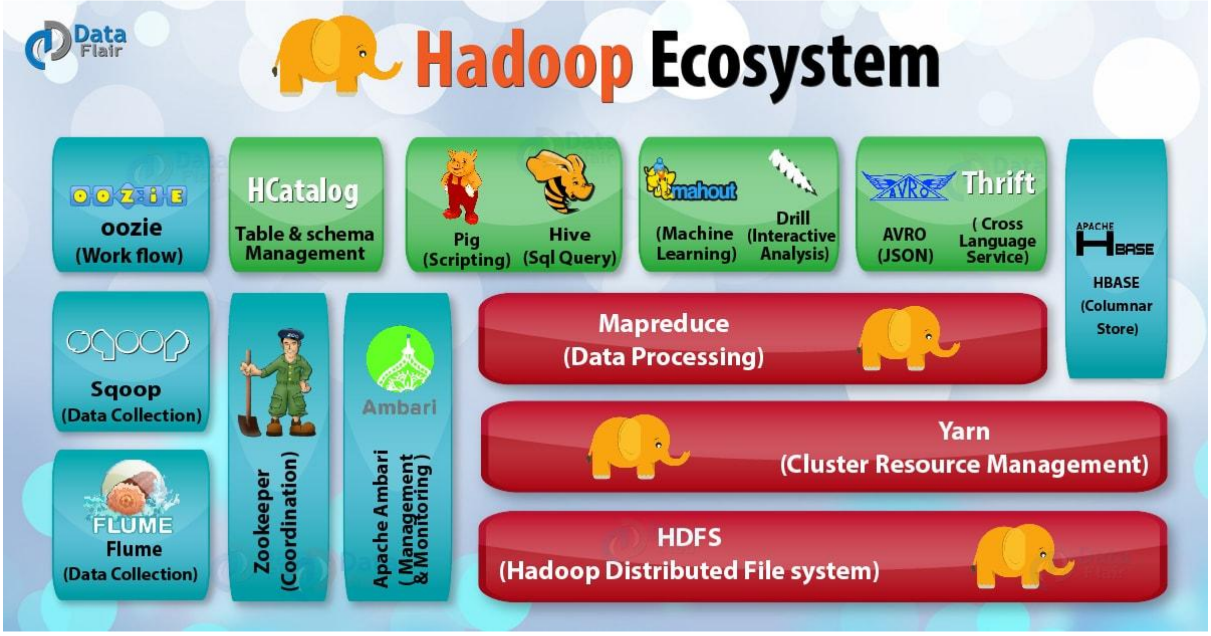
		and JavaScript extensions	

3.3 Data Analysis

3.3.1 MapReduce

The MapReduce[5] programming paradigm can solve almost every problem of distributed and parallel computing, and large scale data-intensive computing. A MapReduce program is composed of a map procedure, which performs filtering and sorting, and a reduce method, which performs a summary operation. The "MapReduce System" (also called "infrastructure" or "framework") orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance.

MapReduce means Apache Hadoop and all the category of tools and solutions that are part of his ecosystems as depicted in the following picture[6].



Technology	License	Description	Origin

[5] https://en.wikipedia.org/wiki/MapReduce

[6] https://data-flair.training/blogs/hadoop-ecosystem-components/

Hadoop	Apache License 2.0	The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.	Global
Cloudera	Apache Software License (ASL) and the Affero General Public License (AGPL)	Cloudera's CDH platform is the most popular distribution of Hadoop and related projects. Since the beginning of 2020 it stopped providing a free version of the distribution.	US

3.3.2 Apache Spark

Apache Spark is a lightning-fast cluster computing technology, designed for fast computation. It is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more types of computations, which includes interactive queries and stream processing. The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application.

Out of the box, Spark can run in a standalone cluster mode that simply requires the Apache Spark framework and a JVM on each machine in your cluster. However, it’s more likely to take advantage of a more robust resource or cluster management system to take care of allocating workers on demand. In the enterprise, this will normally mean running on Hadoop Yarn, but Apache Spark can also run on Apache Mesos, Kubernetes, and Docker Swarm. As a managed solution, then Apache Spark can be found as part of Amazon EMR, Google Cloud Dataproc, Microsoft Azure HDInsight and Databricks,

At the heart of Apache Spark is the concept of the Resilient Distributed Dataset (RDD), a programming abstraction that represents an immutable collection of objects that can be split across a computing cluster. Operations on the RDDs can also be split across the cluster and executed in a parallel batch process, leading to fast and scalable parallel processing. RDDs can be created from simple text files, SQL databases, NoSQL stores (such as Cassandra and MongoDB), Amazon S3 buckets, and much more besides. Much of the Spark Core API is built on this RDD concept, enabling traditional map and reduce functionality, but also providing built-in support for joining data sets, filtering, sampling, and aggregation.

3.3.3 BSL (Bulk Synchronous Parallel)

Bulk-synchronous parallelism[7] is a type of coarse-grain parallelism, where inter-processor communication follows the discipline of strict barrier synchronization. Depending on the context, BSP can be regarded as a computation model for the design and analysis of parallel algorithms, or a programming model for the development of parallel software. A BSP algorithm proceeds in a series of global *supersteps*, which consists of three components:

- *Concurrent computation*: every participating processor may perform local computations, i.e., each process can only make use of values stored in the local fast memory of the processor. The computations occur asynchronously of all the others but may overlap with communication.
- *Communication*: The processes exchange data between themselves to facilitate remote data storage capabilities.
- *Barrier synchronisation*: When a process reaches this point (the *barrier*), it waits until all other processes have reached the same barrier.

The computation and communication actions do not have to be ordered in time. Communication typically takes the form of the one-sided *put* and *get* Direct Remote Memory Access calls, rather than paired two-sided *send* and *receive* message passing calls. The barrier synchronization concludes the superstep: it ensures that all one-sided communications are properly concluded. Systems based on two-sided communication include this synchronisation cost implicitly for every message sent.

Technology	License	Description	Origin
Apache Hama	Apache License 2.0	Framework for Big Data analytics which uses the Bulk Synchronous Parallel (BSP) computing model	Global
Pregel	Proprietary	Data flow paradigm and system of or large-scale graph processing created at Google to solve problems that are hard or expensive to solve using only the MapReduce framework. While the system remains proprietary at Google, the computational paradigm was adopted by many graph-processing systems, and many popular graph algorithms have been converted to the Pregel[8] framework. Spark Graphx and ArangoDB provide support for Pregel API.	US

[7] https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-09766-4_311

[8] https://blog.acolyer.org/2015/05/26/pregel-a-system-for-large-scale-graph-processing/

Apache Giraph	Apache License 2.0	Iterative graph processing system built for high scalability. Giraph originated as the open-source counterpart to Pregel, adding several features beyond the basic model, including master computation, sharded aggregators, edge-oriented input, out-of-core computation, and more.it is currently used at Facebook to analyze the social graph formed by users and their connections.	Global
BSPLib	MIT license	Fast, and easy to use C++ implementation of the Bulk Synchronous Parallel (BSP) threading model. This model is mainly used in the scientific computing field, but can also be applied more generally in computer science.	US

3.3.4 Stream processing

Stream processing is the act of continuously incorporating new data to compute a result. In stream processing, the input data is unbounded and has no predetermined beginning or end. It simply forms a series of events that arrives at the stream processing system e.g. credit card transactions, clicks on a website, or sensor readings from internet of things devices.

Technology	License	Description	Origin
Apache Spark Streaming	Apache License 2.0	Spark Streaming is an extension of the core Spark API that enables scalable, high-throughput, fault-tolerant stream processing of live data streams.	Global
Apache Storm	Apache License 2.0	Distributed realtime computation system. It processes unbounded streams of data, doing for real time processing what Hadoop did for batch processing.	Global
Apache Samza	Apache License 2.0	Allows to build stateful applications that process data in real-time from multiple sources including Apache Kafka.	Global
Apache Flink	Apache License 2.0	Framework and distributed processing engine for stateful computations over unbounded and bounded data streams. Flink has been designed to run in all common cluster environments,	Global

		perform computations at in-memory speed and at any scale.	
Apache Beam	Apache License 2.0	Unified programming model to define and execute data processing pipelines, including ETL, batch and stream (continuous) processing Beam Pipelines are defined using one of the provided SDKs and executed in one of the Beam's supported runners (distributed processing back-ends) including Apache Flink, Apache Samza, Apache Spark, and Google Cloud Dataflow	Global
Amazon Kinesis	see in section 3.1		
Google DataFlow	Proprietary	Data processing service for both batch and real-time data streaming applications. It enables developers to set up processing pipelines for integrating, preparing and analyzing large data sets.	US
MS Azure Stream Analytics	Proprietary	Same data processing service as DataFlow but implemented by Microsoft.	US
IBM Streams	Proprietary	Same data processing service as DataFlow but implemented by IBM.	US
Software AG APAMA	Proprietary & Community Edition	Allows real time processing of high volume of data. Built on in-memory architecture which can be run on a local server or cloud platform. It supports also standard protocol (MQTT and AMQP) for better integration with the IoT world	US

3.4 Data Visualization

Data Visualization allows interaction with the data, giving life and meaning to the data analysis, and keeping users interested in the information. Data Visualization or better referred as "Visual Analytics," uses higher-level tools and even programming tools and libraries that support it.

Technology	License	Description	Origin
Jupyter	3-Clause BSD	The Jupyter Notebook is an open-source web application that allows users to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning. Leverage big data tools, such as Apache Spark, from Python, R and Scala.	Global
R-Shiny	GPL-3	Shiny is an R package that makes it easy to build interactive web apps straight from R. It can be used to host standalone apps on a webpage or embed them in R Markdown documents or build dashboards.	Global
Tableau	Proprietary	Tableau desktop is a data visualisation tool (SaaS) for manipulating big data. It has two other variants "Tableau Server" and cloud-based "Tableau Online" which are dedicatedly designed for big data-related organisations. it can connect data from a spreadsheet to as big as Hadoop and analyse it. Tableau Desktop is free for students and instructors	US
D3	BSD-3-Clause License	D3 or Data Driven Document is a Javascript library for big data visualisations. This is not a tool so users need a good knowledge over javascript to give the collected data a shape. D3 supports large data sets in real-time and it also produces dynamic interaction and animation in both 2D and 3D.	Global
Microsoft Power BI	Proprietary	Microsoft Power BI is a suite of business analytics tools primarily meant for analyzing data and sharing the	US

		insights. It helps deriving quick answers from the data and also can connect to on-premises data sources for real time mapping and analysis. For basic requirements, the tool can even be used for free and it let analyse up to 1GB of data per user account without a paid subscription	
Oracle Visual Analyzer	Proprietary	This web-based tool within the Oracle Business Intelligence Cloud Service provides interactive visuals and advanced analysis with a customisable dashboard.	US

4 Analysis on the current adoption and usage of the technology

4.1 Big Data Collection/Ingestion

When dealing with BigData collection or ingestion the main player in the case of event streaming is Apache Kafka, which is becoming a de-facto standard. Many companies are using Kafka because it provides the functionality of a messaging system but with a distributed architecture which offers durability and fault-tolerance. The usage of more standard messaging systems, like RabbitMQ and ActiveMQ, is quite rising in the IoT world because of the support for MQTT protocol, which is becoming the standard for IoT messaging.

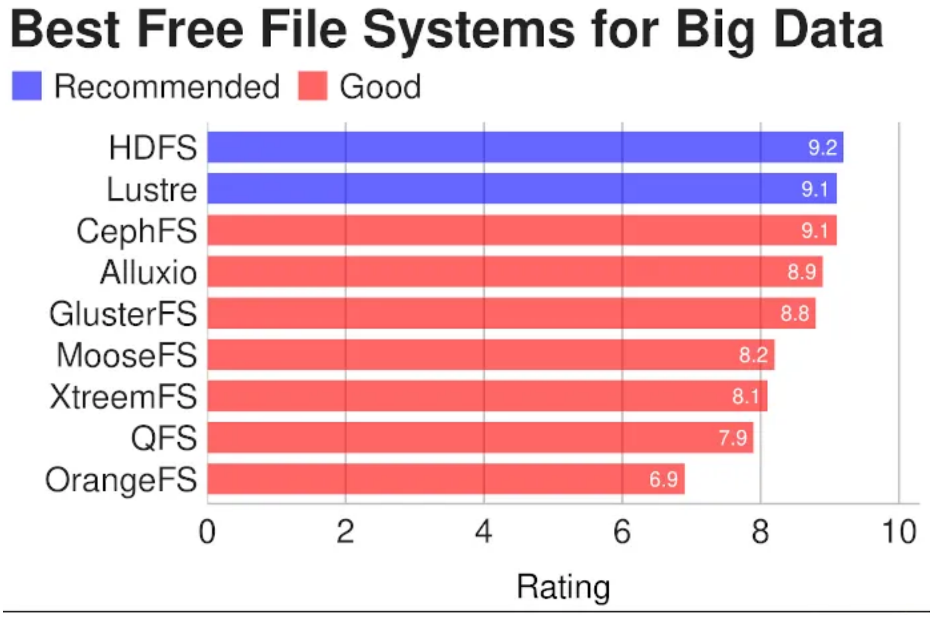
For bulk data transfer between structured datastores such as relational databases and Apache Hadoop HDFS, tools like Apache Sqoop are quite used, while for logs and file ingestions in general Apache Flume is the solution mainly adopted.

In the commercial world, all the various public cloud vendors implemented services for data collections and ingestion, therefore their usage depends on the popularity of the solution for Big Data Analysis then chosen.

4.2 Big Data Storage

We reported different categories of storage solutions than can be used for Big Data, and they all depend on the architecture and the use cases supported.

In the category for file systems for Big Data, we can refer to the analysis published by LinuxLinks[9] comparing free file systems for Big Data and depicted in the following picture.



The filesystems that have been included also in our list have of course a distributed nature, which is fundamental to scale to the size and throughput needed by Big Data tools.

In particular if we take into account Big Data using the Map-reduce programming model, HDFS still provides the best performance and it's the preferred choice. CephFS is also becoming a quite known solution, also after the RedHat effort and acquisitions in the year which lead it to be the major contributor to Ceph.

In the scenario of the EGI infrastructure, several distributed storage solutions are available which have been developed in the "grid" era and in particular for the High Energy physics communities, namely DPM, dCache, StoRM and EOS . New solutions are getting commissioned on the sites of the infrastructure, based also on the already cited solutions, for instance ECHO which is based on CEPH

[9] https://www.linuxlinks.com/FileSystems/

or the OSG HDFS SE where both CEPH and HDFS have been extended with plugins to enable external access via gridftp and Xrootd protocols.

For what concerns the databases, we have described 5 types mainly: relational, key-value store, NoSQL columns based, NoSQL document based and graph databases.

Relational databases can easily handle volume of data in Big Data scenarios, but they are not so appropriate due to the need of a schema definition and for the speed of the ingestion of data (this is not comparable to NoSQL mainly due to the constraints and indexes on the tables). Recently though the support for JSON data types (available in MySQL, PostgreSQL and SQL server) brought into the light the usage of relational databases also in some of Big Data scenarios. It’s evident though that the new generation NoSQL and graph databases are more suitable for the Big Data workloads and we are going to analyze their usage and adoption.

First of all among the NoSQL databases, some reports[10] positions the key-value store as the ones with largest market share, followed by document, column and graph based.

For NoSQL key-value based databases, the commercial solutions like Amazon DynamoDB or Oracle NOSQL database are the preferred ones, with the Oracle DB having also a community edition. Redis and Riak KV are also very popular, together with Aerospike (which has also a community edition).

Among the NoSQL document based solutions, MongoDB is the most common one, available as community edition or Cloud-hosted. The other solutions, with less adoption, also available on the market are CouchDB and RethinkDB.

On the front of the NoSQL column based databases, Cassandra is the most known open source solution followed by HBase and Hypertable, while on the commercial side Google BigTables is the preferred one. Lastly for graph databases, Neo4J community edition and ArangoDB are quite used open source solutions followed by OrientDB, while on the commercial side Amazon Neptune is the leader.

4.3 Big Data Analysis

In the area of Big Data Analysis we have mentioned 4 different sub areas based on the programming models: MapReduce, Apache Spark, BPL and Stream processing.

MapReduce means Hadoop, which, besides the open source project, is distributed by many companies that packaged the hadoop components and forks, like Cloudera which is the most known and used, Hortonworks recently acquired by Cloudera, MapR now part of HP. As described in section 3, Apache Spark, originally developed as part of the Hadoop ecosystem, is now becoming quite independent and it can be executed without YARN and HDFS. According to a survey that Unravel Data [11] conducted in January 2019, Spark was the second most deployed big data technology, with 31% of

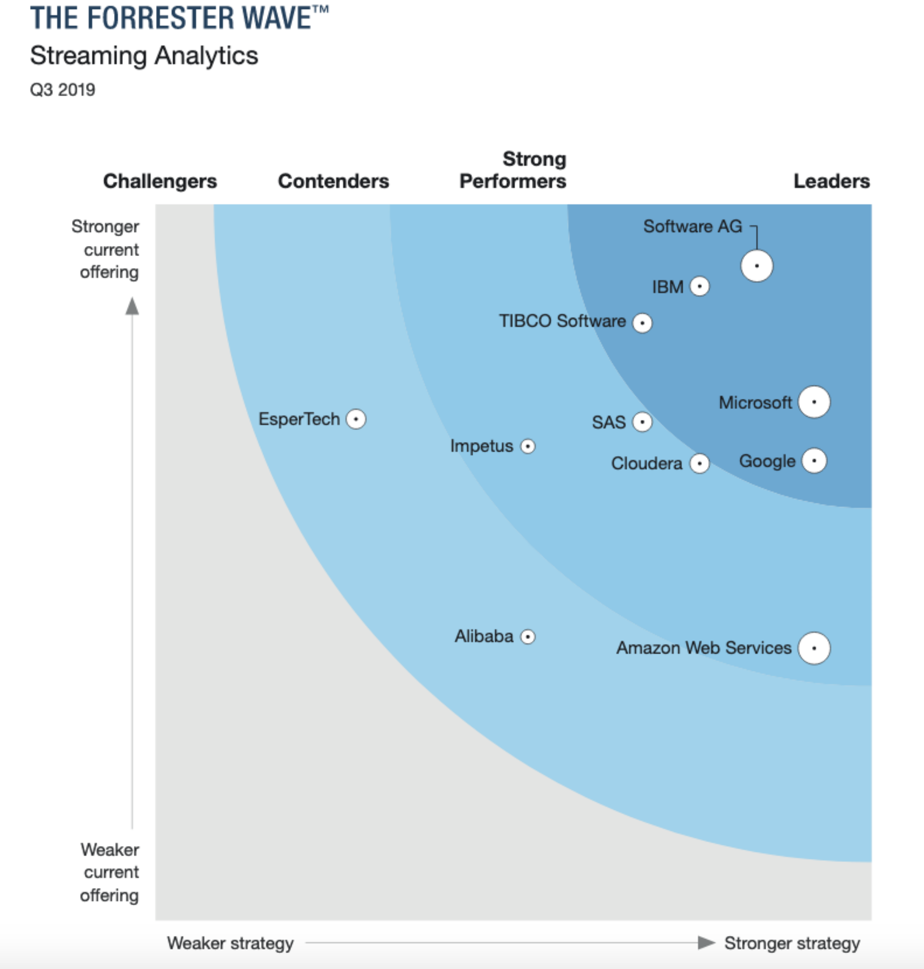
[10] https://www.alliedmarketresearch.com/NoSQL-market
[11] https://unraveldata.com/big-data-survey-2019/

respondents deploying it, compared to 32% for Hadoop. That’s evidence that Spark is on the way to transcending Hadoop. Spark technology has been a huge success, and become a critical component of many big data projects at companies and other organizations around the world. In fact, Spark has been such a massive hit that Databricks, the private Spark-as-a-service company founded by the creators of Apache Spark, is valued at $ Billions.

The BSP programming model is also used in Big Data scenarios and interest has soared in recent years, with Google adopting it as a major technology for graph analytics at massive scale via technologies like Pregel and MapReduce. Also, with the next generation of Hadoop, decoupling the MapReduce model from the rest of the Hadoop infrastructure, there are now active open source projects to add explicit BSP programming (like Spark Graphx), as well as other high performance parallel programming models, on top of Hadoop. The technology mainly known is Apache Giraph, currently used at Facebook to analyze the social graph formed by users and their connection.

Without any doubt stream processing is becoming essential as the adoption of artificial intelligence in the enterprise hits a decisive moment. Artificial Intelligence (AI) and real-time data have formed a closely interlinked duo. Both bring massive technological advances to the enterprise with AI relying on the availability of data to feed models with information, in real time or with low latency. Streaming data platforms provide the fabric behind the scenes to deliver data to machine learning models. They also forward insight to users in real time so that they can generate immediate value out of the data. In addition to AI, the boom of the IoT device is also pushing for stream processing. Today, stream processing is used in every industry and government bodies where stream data is generated through IoT data, for instance in Smart Cities where smart traffic lights, using traffic volume data, bus stops with digital arrival time boards, smart meter reading systems, smart intersection systems and intelligent street lighting.

The following picture describes the Forrester Wave for Streaming Analytics leading companies.



Concerning the Open Source world, the main solutions are divided into micro-batch vs streaming support.

Micro-batch processing is useful when we need very fresh data, but not necessarily real-time, meaning we can't wait an hour or a day for a batch processing to run, but we also don't need to know what happened in the last few seconds. Example scenarios could include web analytics (clickstream) or user behavior. If a large ecommerce site makes a major change to its user interface, analysts would want to know how this affected purchasing behavior almost immediately because a drop in conversion rates could translate into significant revenue losses. However, while a day's delay is definitely too long in this case, a minute's delay should not be an issue making micro-batch processing a good choice. Apache Spark Streaming is the most popular open-source framework for micro-batch processing.

Stream processing is used instead when we need to analyze or serve data as close as possible to when we get hold of it. Examples of scenarios where data freshness is super-important could include real-time advertising, online inference in machine learning, or fraud detection. In these cases we have data-driven systems that need to make a split-second decision and the use of stream processing is fundamental to quickly access the data, perform our calculations and reach a result. Common solutions are Apache Flink, Apache Samza and Apache Storm.

4.4 Big Data Visualization

The tools for Big Data visualization should provide a certain set of features:

- Capability to apply various filters to adjust the results
- Capability to interact with the data sets during the analysis
- Capability to connect to other software to receive incoming data or provide input for them
- Capability to provide collaboration options for the users

For sure the most known solution nowadays is Jupyter notebook, because it combines Big Data analysis with visualization and the possibility to share amongst the team to enable internal collaboration on the data analysis. While being initially using Python and R, Jupyter Notebook is actively introducing kernels for other programming languages like Java, Go, C#, Ruby, and many others. When using R, Shiny is also a quite used solution to build interactive web applications. Libraries are also available like Google Chart or D3.js

On the commercial side Tableau is the market leader, especially efficient for delivering interactive data visualization for the results derived from Big Data operations, deep learning algorithms and multiple types of AI-driven apps. Microsoft Power BI and Oracle Data visualizer are other examples of known commercial productis.

5 Standardisation activities & policies

The following standards have been quite recently published:

- ISO/IEC 20546 Big data - Overview and vocabulary
- ISO/IEC 20547 Information technology - Big data reference architecture
    - Framework and application process
    - Use cases and derived requirements
    - Reference architecture
    - Security and privacy
    - Standards roadmap

More details at
https://etech.iec.ch/issue/2020-02/new-iec-and-iso-standard-will-enable-big-data-adoption-across-industry-sectors

Other relevant documents are the *EU policy in the context of BigData*[12] and theData Governance Act [13].

6 Relevant partners in the field

This section lists technology providers collaborating with EGI, members of the EGI Federation and other infrastructures that are working on solutions for Big Data.

6.1 Major technology providers

Partner	Expertise	Tools
Atos Codex	Atos is partner with EGI in some EU funded projects and they provide also Big Data solutions	https://atos.net/en/solutions/atos-codex-connected-intelligence

6.2 Interested partners of the EGI Federation

Partner	Expertise	Tools
UPV	IM components, deployment of clusters (hadoop) and EC3 for scaling	https://www.grycap.upv.es/im/index.php http://servproject.i3m.upv.es/ec3/
CERN	CERN has developed since many years the tools for storage, analytics that have been reused by many EGI communities. Lately the	https://swan.web.cern.ch/swan/

[12] https://ec.europa.eu/digital-single-market/en/policies/big-data

[13] https://ec.europa.eu/digital-single-market/en/news/proposal-regulation-european-data-governance-data-governance-act

	development of the SWAN service, which is provided also as on-premises solution could be seen as general solution for some EGI communities	
BI Insight	https://biinsight.pl/en Pilots in EOSC DIH https://eosc-dih.eu/access-the-knowledge/	
Odin solutions	https://www.odins.es Pilots is EOSC DIH https://eosc-dih.eu/bigcoldtruck/	
INFN	DODAS (Dynamic On Demand Analysis Service) provides the end-user with an automated system that simplifies the process of provisioning, creating, managing and accessing a pool of heterogeneous (possibly opportunistic) computing resources. DODAS allows to generate both HTCondor batch systems and BigData platforms such as Spark, HDFS with pluggable multi-cloud support	DODAS https://dodas-ts.github.io/dodas-doc

6.3 Involvement of other e-infrastructures

e-Infrastructure/ Partner	Expertise	Tools
SoBigData	SoBigData is the European Research Infrastructure for Big Data and Social Mining. They have recently applied to to become a research infrastructure recognized by ESFRI RoadMap 2021. EGI is also a partner	Tools for data mining and big data analytics integrated in the D4science infrastructure

7 Projects, Initiatives, Communities and partnerships

7.1 BDVA - https://www.bdva.eu

The Big Data Value Association (BDVA) is an industry-driven international not–for-profit organisation with more than 200 members all over Europe and a well-balanced composition of large, small, and medium-sized industries as well as research and user organizations. BDVA is the private counterpart to the EU Commission to implement the Big Data Value PPP program. BDVA and the Big Data Value PPP pursue a common shared vision of positioning Europe as the world leader in the creation of Big Data Value. The mission of the BDVA is to develop the Innovation Ecosystem that will enable the data and AI-driven digital transformation in Europe delivering maximum economic and societal benefit, and, achieving and sustaining Europe’s leadership on Big Data Value creation and Artificial Intelligence. EGI is one of the BDVA iSpaces.

7.2 EUHubs4Data - https://euhubs4data.eu

The European federation of Data Driven Innovation Hubs aims to consolidate as the European reference for data driven innovation and experimentation, fostering collaboration between data driven initiatives in Europe, federating solutions in a global common catalogue of data services, and sharing data in a cross-border and cross-sector basis. With the objective of serving as reference to the establishment of the Common European Data Spaces, the federation is initially composed of 12 DIHs, covering 10 countries and 12 different regions, and plans to increase the geographical coverage by incorporating other relevant initiatives in the future. The project started in september 2020 and has a duration of 3 years. EGI is WP leader for the building of the federated data catalogue.

7.3 BD4NRG - Big Data for Energy

The project is starting in January 2021 with a duration of 3 years, and has as objectives:

- deliver a reference architecture for Smart Energy, which aligns BDVA SRIA, IDSA and FIWARE architectures, SAREF standard and extend COSMAG specification to enable B2B multi-party data exchange, while providing full interoperability of leading-edge big data technologies with smart grid standards and operational frameworks
- evolve and upscale a number of TRL 5-6 technology enablers, such as scalable sovereignty preserving hybrid DLT/off-chain data governance, big data elastic pipeline orchestration, IoT/edge AI-based federated learning and multi-resource sharing tokenized marketplace, loosely integrate and deploy them within the TRL 7-8 BD4NRG framework
- deliver a TRL8 open modular big data analytic toolbox as front-end for one-stop-shop analytics services development by orchestrating legacy and/or third party assets (data, computing resources, models, algorithms)

- validate such framework through the delivery of predictive and prescriptive edge AI-based big data analytics on 13 large scale pilots, deployed by different energy stakeholders (TSOs and DSOs power network operators, aggregators, storage/renewable assets operators, local energy communities, ESCOs, power market operators, municipalities, financial institutions and ENTSO-E), fully covering the energy value chain
- setup a vibrant data-driven ecosystem through the SGBDAA Alliance, which will federate new energy data providers, attract SMEs for novel energy services provisioning through cascading funding and validate a hybrid energy/industry value chain supporting B2B joint digital platforms.

7.4 SoBigData++

SoBigData++ strives to deliver a distributed, Pan-European, multi-disciplinary research infrastructure for big social data analytics, coupled with the consolidation of a cross-disciplinary European research community, aimed at using social mining and big data to understand the complexity of our contemporary, globally-interconnected society EGI is a e-infrastructure engineer partner. EGI will enhance the SoBigData platform with two services: Jupyter Notebooks and the Workflow manager Galaxy. The project started in January 2020 and it has a duration of 4 years.

7.5 RDA

At least one Interest Group (IG) in RDA is dealing with Big Data and Big Data Analytics:

https://www.rd-alliance.org/groups/big-data-analytics-ig.html

It has already delivered some reports, eg. Big Data storage and data virtualization:
https://www.rd-alliance.org/system/files/documents/ajit%20Paper%20-%20bigdata%20storage%20and%20data%20virtualization.doc

8 Integration scenarios in the EGI Infrastructure

Possible Integration scenario in the EGI infrastructure can vary depending on the use case.

8.1 Access to Big Data Analytics tools

EGI is already providing for some communities tools for Data Analytics that have been developed mainly in the context of the HEP communities (ie. LHC experiments) and that are also included or planned to be included in the EGI portfolio:

- Frontier Squid and CVMFS for the distributions of reference datasets (i.e. Calibrations) and software to the computing infrastructure and optimized caching
- Pre-stage of datasets to computing nodes, integrated with Workflow management systems and experiments frameworks (DIRAC, Rucio, FTS, etc)
- Optimized remote files access (via data federations)

But also in the context of LHC, there are many projects to align the tools to the one used by industry (Hadoop, Spark) and CERN has already since some years provided clusters and services for this purpose[14], like the Hadoop deployment made available by the IT department[15] and the SWAN service[16] which combines the Jupyter with storage and Spark cluster access.

As already described in the case of CERN, the natural way to support Data Scientist would be to provide at one or more the EGI Federation site, some of the tools that has been described in the document. As a simple example, the possibility to access Spark clusters from Jupyter notebooks, (like in SWAN) which is becoming quite a popular solution, in conjunction with easy access to data co-located at the same site. So for instance one possible integration would be to deploy Spark cluster colocated with EGI notebooks solution which we already provide.

8.2 Automatic deployment of Big Data analytics tools and auto-scaling

There are several tools already available that can be used to automatically deploy Big Data tools on the infrastructure (i.e. the Infrastructure Manager) and also provide the automatic scaling of resources according to users needs. This is a slight different use case compared to the previous one, as this is mainly for users that would like to access resources not already pre-deployed and available in a “shared” environment, but more a dedicated installations that can also be autoscaled if needed and destroyed in the case they are not used.

As said we already provide the Infrastructure Manager service and EC3 to automatic the installation and scaling of resources, both provided by UPV, but also the DODAS service by INFN falls in this area (new service to be integrated in EGI-ACE). They could be extended to automatically provision more emerging technologies as they become popular.

Together with the automatic deployment, these tools could be also already linked with existing storage solutions for Big-Data available at sites (as reported in section 3). For example deploying

[14] https://www.researchgate.net/publication/321232977_Hadoop_and_friends_-_first_experience_at_CERN_with_a_new_platform_for_high_throughput_analysis_steps

[15] https://www.researchgate.net/publication/335864513_Evolution_of_the_Hadoop_Platform_and_Ecosystem_for_High_Energy_Physics

[16] https://swan.web.cern.ch/swan/

automatically Spark clusters at sites which provide already HDFS storage, this of course requires knowledge on the Storage solutions available at the different sites part of the EGI federation.

8.3 Collection and storage of IoT data and stream processing

Another possible scenario is to offer services for the collection of data coming from IoT devices in conjunction with stream processing and storage. For what concerns the data collection and ingestion there are several tools that have been described in section 3.1 and could be made available as a service in the Federation (Active brokers, Kafka, etc)

The deployment of Stream processing services could also be envisaged together with storage and data visualization tools described in sections 3.2, 3.3 and 3.4.