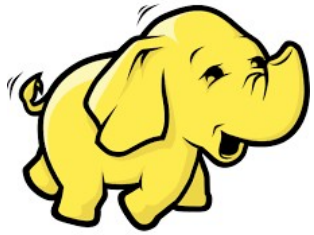
Hadoop vs Others

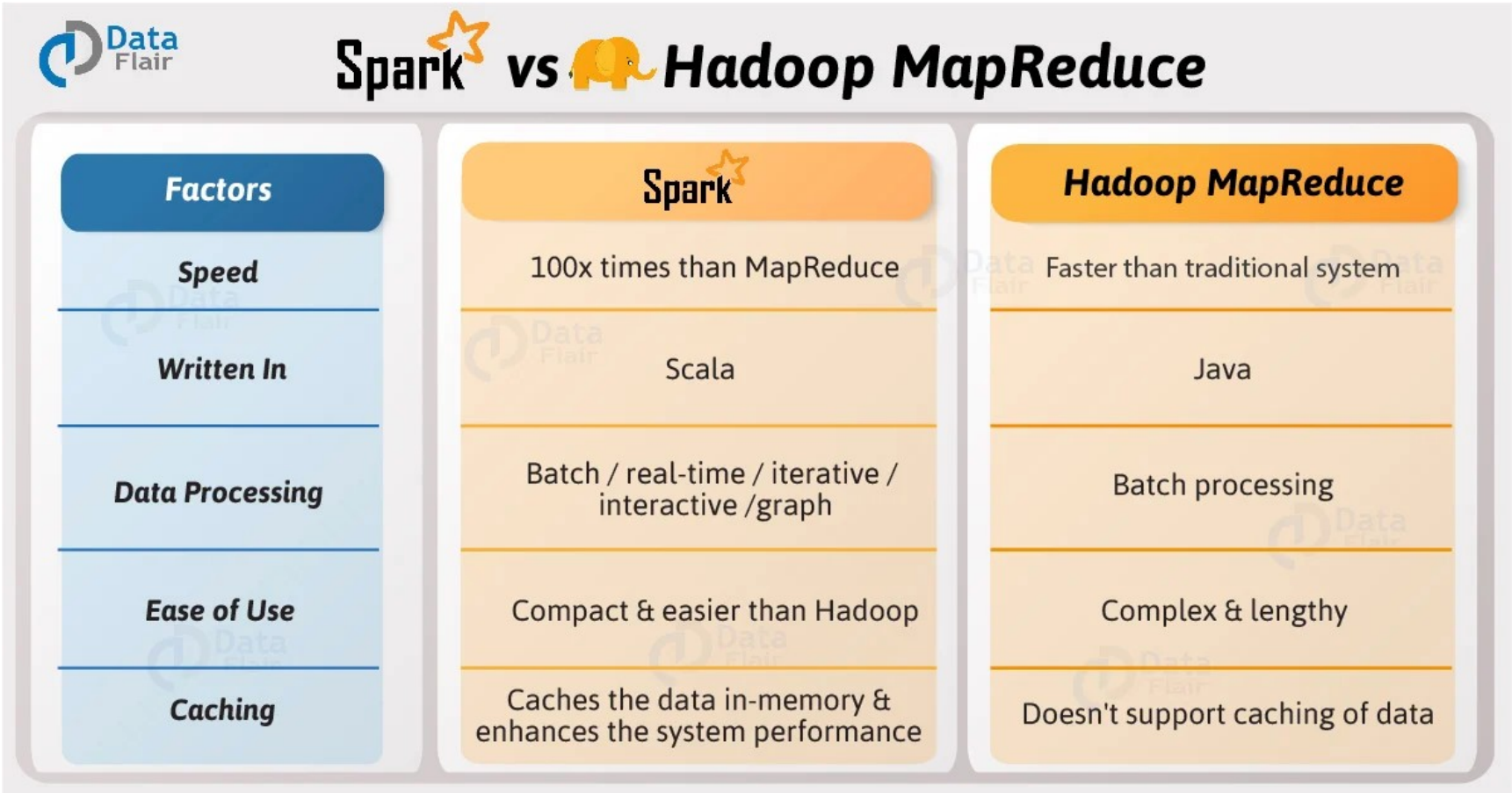
Tushar B. Kute,
http://tusharkute.com



Apache Spark™ and Apache™ Hadoop®

- Spark is an open source, in-memory, parallel data processing engine.
- It is an alternative to MapReduce which is less widely used these days. Spark can run on Hadoop or on its own cluster.

Apache Spark™ and Apache™ Hadoop®



Big data and Hadoop

- Big data is data with the 3 Vs
  - Volume (too big to be handled by conventional database technologies)
  - Velocity (it changes quickly)
  - Variety (the data contains many different types of data – both structured and unstructured)
- Hadoop is the open source technology that allows big data to be stored and processed on low cost industry standard servers.
- Prior to Hadoop, storing and processing big data on conventional technology was very expensive. Hadoop was seen as a way to solve the problem.

Cassandra™ and Hadoop

- Cassandra is a NoSQL distributed database which runs on clusters of commodity servers.
- It is designed to cope with large amounts of data and support high availability by having no single point of failure.
- It is designed for de-normalized data and has its own programming language (CQL). It is therefore more suited for high throughput production processes rather than ad-hoc, self-service data analytics.
- Hadoop, on the other hand, is a general purpose big data storage and data processing platform. It also runs on clusters of low cost commodity hardware.

Cassandra™ and Hadoop

- A plethora of additional “Hadoop applications” allow Hadoop clusters to perform a wide variety of data related tasks.
- Cassandra can run on Hadoop and so can be considered as a Hadoop application. Both Hadoop and Cassandra are open source software.

Elasticsearch™ and Hadoop

- Elasticsearch is a big data search engine that is optimized for text searching.
- It runs on clusters of commodity servers and is fault tolerant. Typical use cases for Elasticsearch are log and click stream analysis.
- Hadoop, on the other hand, is an open source, general purpose big data storage and data processing platform.
- It also runs on clusters of low cost commodity hardware. A plethora of additional “Hadoop applications” allow Hadoop clusters to perform a wide variety of data related tasks

Greenplum® and Hadoop

- Greenplum is a massively parallel database based on PostgreSQL®. Greenplum was bought by a company called EMC and became integrated into a product called Pivotal.
- Like Hadoop, Greenplum can store large amounts of data and allow it to be processed and analyzed.
- However, unlike Hadoop, which is an open source, general purpose big data platform, running on low cost, industry standard hardware, Greenplum is primarily a SQL (Structured Query language) engine and data must be accessed and processed using Greenplum's interfaces and processing engine.

Greenplum® and Hadoop

- Hadoop, on the other hand supports a plethora of additional “Hadoop applications” allowing Hadoop clusters to perform a wide variety of data related tasks, including SQL interfaces that are much more performant than the Greenplum SQL interface.
- The Greenplum SQL interface is also available as a Hadoop application called HAWQ.

Apache Impala® and Hadoop

- Apache Hadoop is an open source technology for storing and processing extremely large data sets across hundreds or thousands of computing nodes or servers that operate in parallel using a distributed file system.
- A plethora of additional “Hadoop applications” allow Hadoop clusters to perform a wide variety of data related tasks.
- Impala is a Hadoop application, developed and supplied by Cloudera, that allows data stored in Hadoop to be queried via an industry standard SQL interface.

MongoDB® and Hadoop

- MongoDB is a NoSQL database. The main difference between MongoDB and Hadoop is that MongoDB is a database whereas Hadoop is an ecosystem of technologies that allow clusters of low cost industry standard hardware to perform a wide variety of data related tasks.

SQL and Hadoop

- SQL stands for structured query language. It is a language used to query data.
- Hadoop is an ecosystem of technologies that allow clusters of low cost industry standard hardware to perform a wide variety of data related tasks.
- SQL is used to query data in Hadoop through technologies such as Hive, Impala and Kognitio which run on top of Hadoop.

Netezza and Hadoop

- Netezza®, now owned by IBM®, is a massively parallel database appliance. Like Hadoop, Netezza can store large amounts of data and allow it to be processed and analyzed.
- However unlike Hadoop, which is an open source, general purpose big data platform, running on low cost industry standard hardware, Netezza is primarily a SQL (Structured Query language) engine, running on proprietary hardware and data must be accessed and processed using Netezza's interfaces and processing engine.
- Hadoop, on the other hand supports a plethora of additional "Hadoop application" which allow Hadoop clusters to perform a wide variety of data related tasks, including high performance SQL interfaces.

NoSQL and Hadoop

- NoSQL originally literally meant ‘No SQL’. NoSQL is a way of querying data by writing a program rather than using SQL.
- Certain types of data analytics, especially analytics that use complex mathematical algorithms are difficult to implement in standard SQL, hence the emergence of NoSQL based data analytics.

NoSQL and Hadoop

- However in recent years it has become recognized that SQL still has an important role to play in providing pervasive access to data.
- NoSQL is now often defined as Not Only SQL as many products now support combined SQL and NoSQL functionality.
- You can use a NoSQL database on top of Hadoop, but Hadoop itself is not NoSQL in the same way that it is not a relational database, nor an object database.

Oracle® and Hadoop

- Oracle is a database software designed primarily for processing transactions.
- Hadoop, on the other hand, is an open-source, general purpose, big data storage and data processing platform.
- It runs on clusters of low cost commodity hardware.

RDBMS and Hadoop

- Relational Database Management System (RDBMS) is a traditional database that stores data which is organized or structured into rows and columns and stored in tables.
- Hadoop is not a database. It is an open-source, general purpose, big data storage and data processing platform.
- It runs on clusters of low cost commodity hardware. A plethora of additional “Hadoop applications” allow Hadoop clusters to perform a wide variety of data related tasks. Unlike the RDBMS, the data in Hadoop can also be unstructured.

Will Hadoop replace RDBMS?

- For some types of applications yes, but for many applications a dedicated database product would be more suitable than a general-purpose platform like Hadoop.
- Hadoop’s strength is its ability to scale so that it can handle very large datasets. The majority of database applications do not need this scalability.

HP Vertica™ and Hadoop

- HP Vertica is a columnar database. Vertica stores data in columns rather than rows, which allows for greater data compression. It has a scale-out architecture and can therefore, like Hadoop, be used to store and analyze large data sets.
- However, unlike Hadoop, which is an open source, general purpose big data platform, running on low cost industry standard hardware, Vertica is primarily a SQL (Structured Query language) engine and data must be accessed and processed using Vertica's interfaces and processing engine.

HP Vertica™ and Hadoop

- Hadoop, on the other hand supports a plethora of additional “Hadoop applications” allowing Hadoop clusters to perform a wide variety of data related tasks, including high performance SQL interfaces.
- Vertica generally runs on its own infrastructure, but a version is available that will run on Hadoop. The Hadoop version is different to the standalone version and does not have the same functionality.

Best alternatives to Hadoop

- Big data technologies are constantly evolving and what is best for your business depends on how much data you have, what type of data it is, and how you intend to use that data.
- Hadoop is a very general-purpose platform. If there is one particular use your business needs it for, then it is possible that other products might be better suited to that particular purpose.
- However many people choose Hadoop because it's open-source licensing model and commodity hardware infrastructure makes it appear a low cost way of dealing with big data when compared to the traditional licensing model used by most of the alternative technologies.

Best alternatives to Hadoop

- Alternative big-data technologies include
  - Teradata
  - IBM Netezza
  - HP Vertica
  - Kognitio (deployed standalone)
  - Exasol

Thank you

This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com