Big Data Tools

UNIT 1

- ***Big Data*** is an essential part of almost every organization these days and to get significant results through Big Data Analytics a set of tools is needed at each phase of data processing and analysis.

- ***Big data*** is simply too large and complex data that cannot be dealt with using traditional data processing methods.

- Big Data requires a ***set of tools & techniques*** for analysis to gain insights from it.

- ***"A Good Tool improves the way you work. A Great Tool improves the way you think"***

  – Jeff Duntemann, Co-Founder of Coriolis

- Analyzing & processing Big Data is not an easy task.

- Big Data is one big problem and to deal with it you need a set of great big data tools that will not only solve this problem but also help you in producing substantial results.

What are the best Big Data Tools?

Here is the list of top 10 big data tools –

- Apache Hadoop
- Apache Spark
- Flink
- Apache Storm
- Apache Cassandra
- MongoDB
- Kafka
- Tableau
- RapidMiner
- R Programming

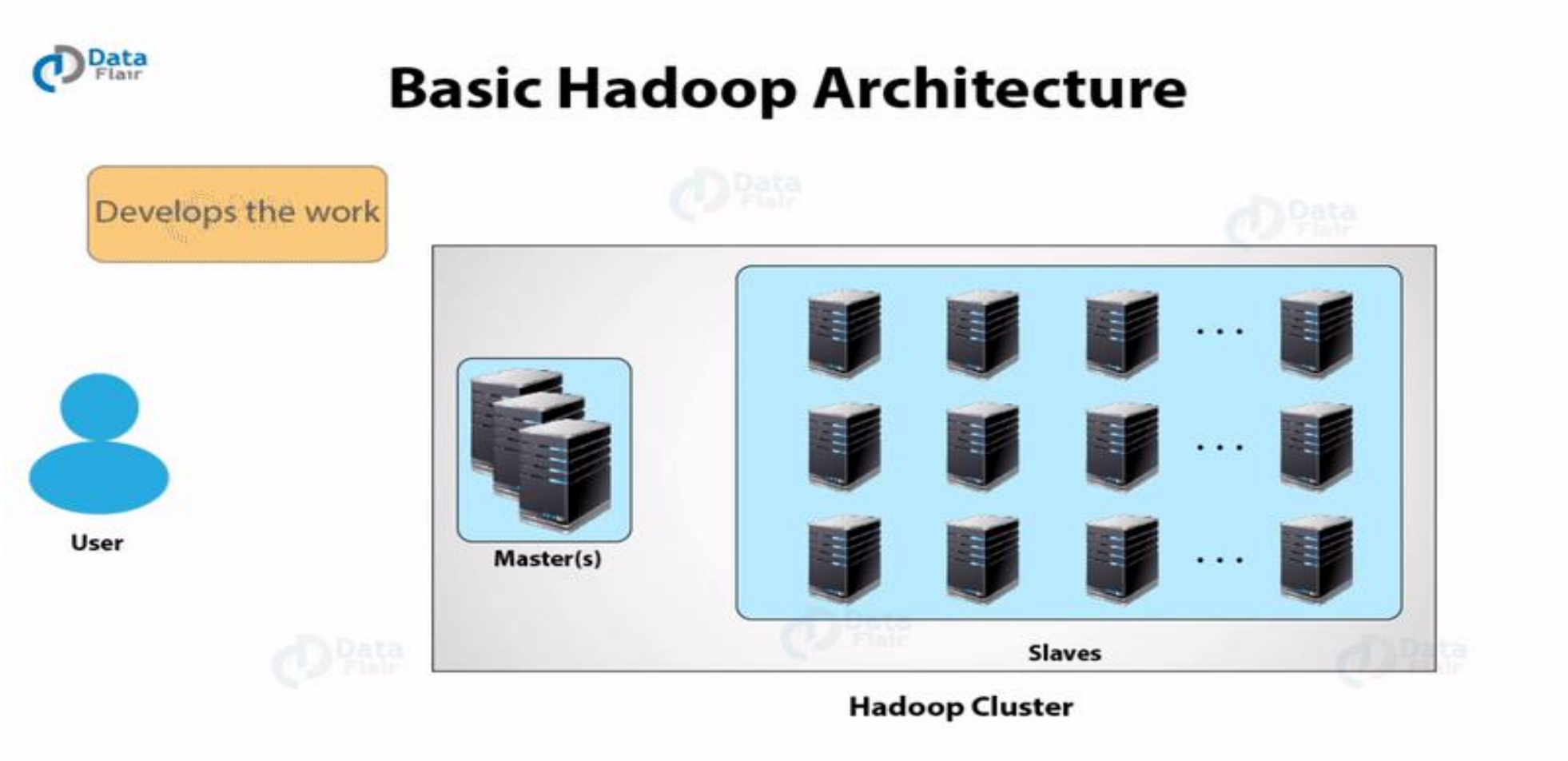
There are a few ***factors*** to be considered while opting for the set of tools i.e., the size of the datasets, pricing of the tool, kind of analysis to be done, and many more.

With the exponential growth of Big Data, the market is also flooded with its ***various tools***. These tools used in big data help in bringing out ***better cost efficiency*** and thus increases the ***speed*** of analysis.

Apache Hadoop

- It is one of the most popularly used tools in the Big Data industry.
- an open-source framework from Apache and runs on commodity hardware. It is used to ***store process*** & ***analyze*** Big Data.
- It is written in Java and enables ***parallel processing*** of data as it works on ***multiple machines*** simultaneously.
- It uses clustered architecture. A Cluster is a group of systems that are connected via LAN.
- Few Disadvantages are:
  - Hadoop does ***not support real-time*** processing. It only supports batch processing.
  - Hadoop ***cannot do in-memory*** calculations.
- It consists of 3 parts-
  - **Hadoop Distributed File System (HDFS) –** It is the ***storage layer*** of Hadoop.
  - **Map-Reduce –** It is the data ***processing layer*** of Hadoop.
  - **YARN –** It is the ***resource management layer*** of Hadoop.

Apache Hadoop



Apache Spark

Apache Spark can be considered as the successor of Hadoop as it overcomes the drawbacks of it.

Spark, unlike Hadoop, supports both ***real-time*** as well as ***batch processing***. It is a general-purpose clustering system.

It also supports ***in-memory*** calculations, which makes it ***100 times*** faster than Hadoop. This is made possible by reducing the number of read/write operations into the disk.

It provides more ***flexibility*** & versatility as compared to Hadoop since it works with different data stores such as HDFS, OpenStack and Apache Cassandra.

It offers high-level APIs in Java, Python, Scala & R.

It also offers a substantial set of ***high-level tools*** including Spark SQL for structured data processing, MLlib for machine learning, GraphX for graph data set processing, and Spark Streaming.

It also consists of ***80 high-level operators*** for efficient query execution.

Apache Storm

It is an open-source big data tool, distributed real-time & ***fault-tolerant*** processing system.

It efficiently processes unbounded streams of data(the data that is ever-growing and has a beginning but no defined end).

It can be used with any of the programming languages and it further ***supports JSON*** based protocols.

The processing ***speed*** of Storm is ***very high***. It is easily scalable and also fault-tolerant. It is much easier to use.

On the other hand, it guarantees the processing of each data set.

Apache Cassandra

It is a distributed database that provides high ***availability*** & ***scalability*** without compromising performance efficiency.

It can ***accommodate*** all types of data sets namely structured, semi-structured, and unstructured.

It is the perfect platform for ***mission-critical data*** with no single point of failure and provides fault tolerance on both commodity hardware and cloud infrastructure.

It works quite ***efficiently*** under heavy loads.

It does ***not*** follow ***master-slave*** architecture so all nodes have the same role.

Apache Cassandra supports the ACID (Atomicity, Consistency, Isolation, & Durability) properties.

Apache Flink

Apache Flink is an Open-source data analytics tool distributed processing framework for bounded & unbounded data streams.

It is written in Java & Scala. It provides ***high accuracy*** results even for late-arriving data.

Flink is a stateful & ***fault-tolerant*** i.e. it has the ability to recover from faults easily.

It provides high-performance efficiency at a large scale, performing on thousands of nodes.

It gives a ***low-latency***, high throughput streaming engine and supports event time & state management.

R Programming

R is an open-source programming language and is one of the most comprehensive statistical analysis languages.

It is a multi-paradigm programming language that offers a dynamic development environment.

As it is an open-source project and thousands of people have contributed to the development of the R.

R is written in C and Fortran. It is one of the most widely used statistical analysis tools as it provides a vast package ecosystem.

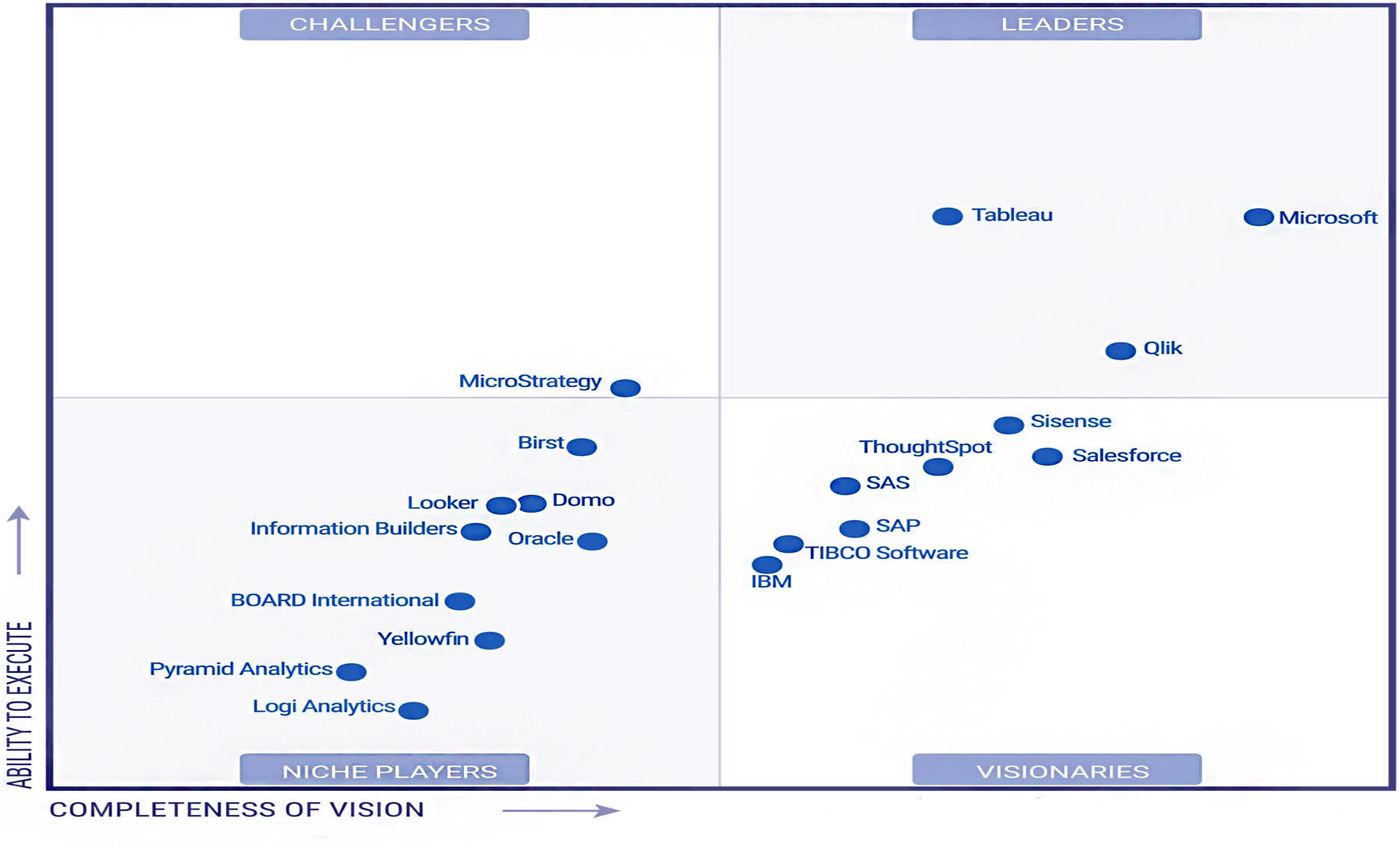
It facilitates the efficient performance of different statistical operations and helps in generating the results of data analysis in graphical as well as text format.

The graphics and charting benefits it provides are unmatchable.

Tableau

- ***Tableau*** is one of the ***best data visualization*** and software solution tools in the BI industry.
- It turns your raw data into valuable insights and enhancing the decision-making process of the businesses.
- It offers a rapid data analysis process and resulted in visualizations are in the form of interactive dashboards and worksheets.
- It works in synchronization with other **Big Data tools** such as Hadoop.
- **Tableau** offered the capabilities of data blending are best in the market.
- It provides an efficient real-time analysis.
- This software doesn't require any technical or programming skills to operate.

Figure 1. Magic Quadrant for Analytics and Business Intelligence Platforms



MongoDB

It is an open-source data analytics tool, NoSQL database that provides ***cross-platform*** capabilities.

It is exemplary for a business that needs ***fast-moving*** & real-time data for taking decisions.

MongoDB is perfect for those who want ***data-driven solutions***.

It is user-friendly as it offers easier installation & maintenance. MongoDB is reliable as well as cost-effective.

It is written in C, C++, and JavaScript.

It is one of the most popular databases for Big Data as it facilitates the management of ***unstructured data*** or the data that changes frequently.

MongoDB uses ***dynamic schemas***. Hence, you can prepare data quickly. This allows in reducing the overall cost. It executes on MEAN software stack, NET applications and, Java platform.

It is also ***flexible*** in cloud infrastructure.

But a certain ***downfall in the processing speed*** has been noticed for some use-cases.

Kafka

Apache Kafka is an open-source platform that was created by LinkedIn in the year 2011.

It is a ***distributed event*** processing or streaming platform which provides high throughput to the systems.

It is *efficient* enough to handle *trillions* of events a day.

It is highly scalable and also provides great ***fault tolerance***.

The streaming process includes publishing & subscribing to streams of records alike to the messaging systems, storing these records durably, and then processing these records. These records are stored in groups called ***topics***.

It offers high-speed streaming and ***guarantees zero downtime***.

RapidMiner

RapidMiner is a cross-platform tool that provides a robust environment for Data Science, ML & Data Analytics procedures.

It is an ***integrated platform*** for the complete Data Science lifecycle starting from data prep to ML to predictive model deployment.

It offers various licenses for small, medium, & large proprietary editions.

It also offers a ***free edition*** that permits only 1 logical processor & up to 10,000 data rows.

RapidMiner is an open-source tool that is written in Java.

It offers ***high efficiency*** even when integrated with APIs and cloud services.