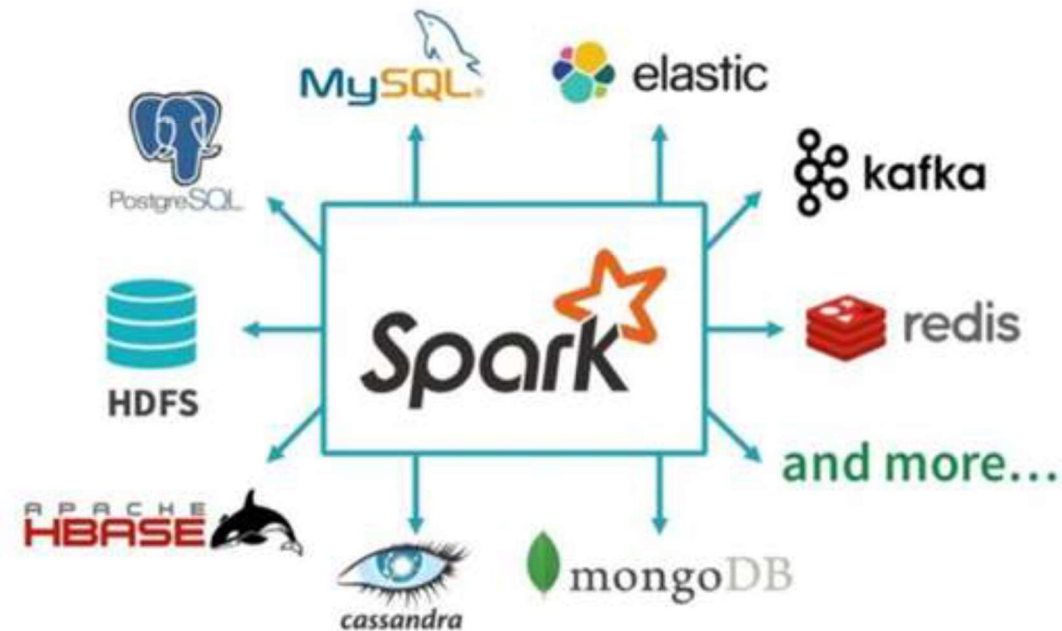
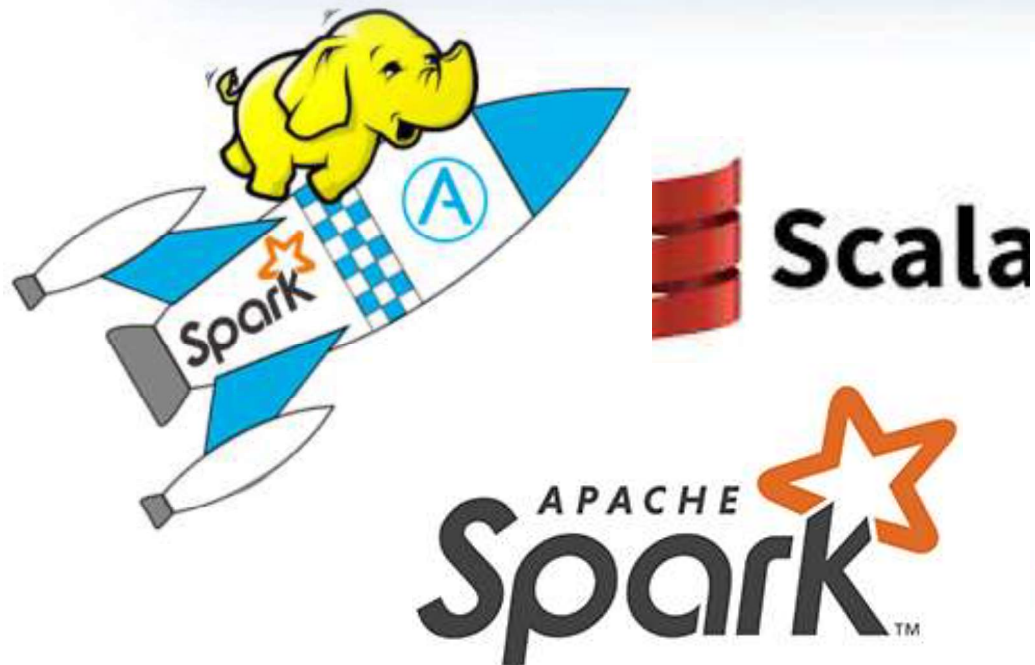


National Institute of Electronics and Information Technology

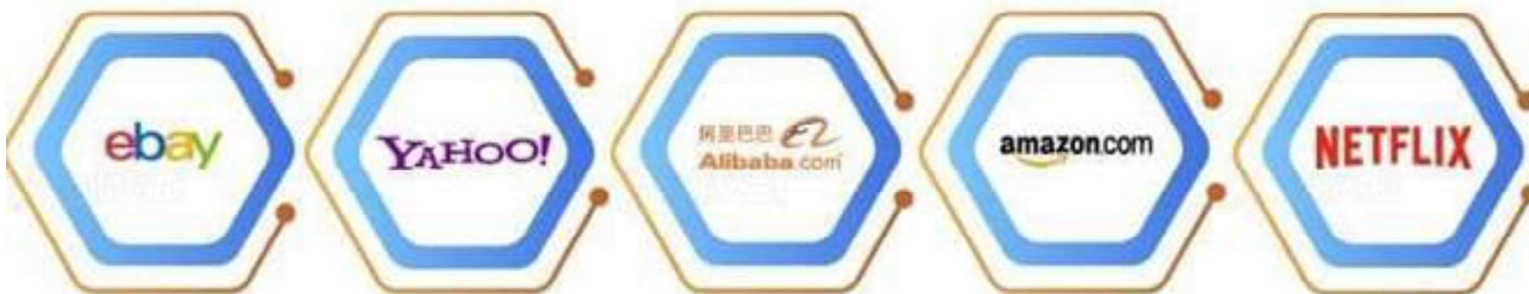
Big Data and Hadoop

Module 7-Ultimate Spark

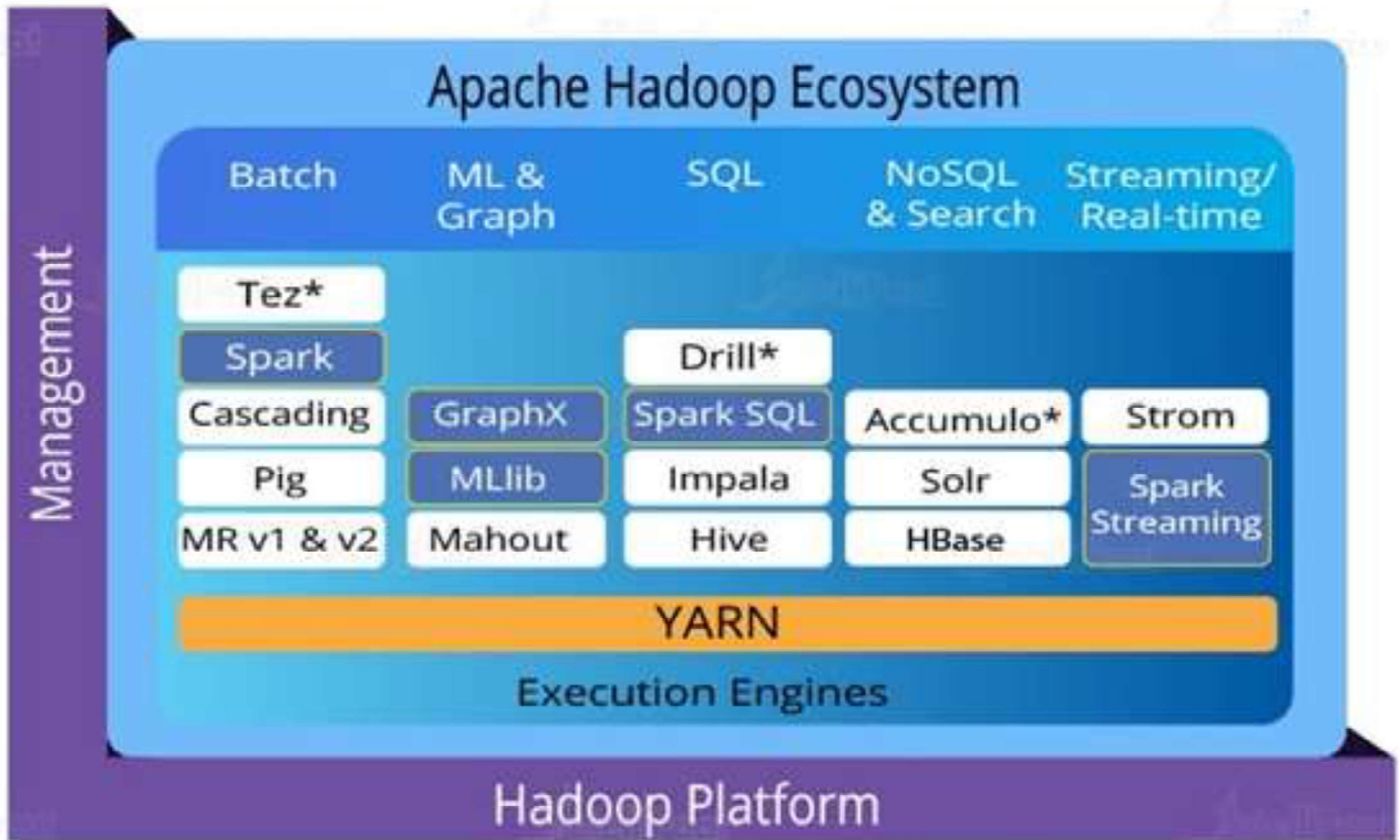


History

1. The Spark got initiated as one of the research projects in 2009 at University of California, Berkeley's AMP Lab. It was later open-sourced in 2010. The major intention behind this project was to create a cluster management framework that supports various computing systems based on clusters.
2. After its release to the market, Spark grew and moved to Apache Software Foundation in 2013. Now, most of the organizations across the world have incorporated Apache Spark for empowering their big data applications.
3. Spark and Hadoop together make a powerful combination to handle Big Data Analytics. Spark can be deployed in numerous ways like in Machine Learning, streaming data, and graph processing.
4. Spark supports programming languages like Python, Scala, Java, and R. In this section, we will understand what Apache Spark is.
5. **Organizations using Spark:**



How does Apache Spark fit in Hadoop ecosystem?



Domain Scenarios of Apache Spark

1. **Banking:** Spark is being more and more adopted by the banking sector. It is mainly used here for financial fraud detection with the help of Spark ML. Banks use Spark to handle credit risk assessment, customer segmentation, and advertising.
2. **E-commerce:** Spark is widely used in the e-commerce industry. Businesses can share their findings with other data sources to provide better recommendations to their customers. Recommendation systems are mostly used in the e-commerce industry to show new trends.
3. **Healthcare:** Apache Spark is a powerful computation engine to perform advanced analytics on patient records. The healthcare industry uses Spark to deploy services to get insights such as patient feedbacks, hospital services, and to keep track of medical data.
4. **Media:** Many gaming companies use Apache Spark for finding patterns from their real-time in-game events. With this, they can derive further business opportunities by customizing such as adjusting the complexity-level of the game automatically according to players' performance, etc.
5. **Travel:** Many people land up with travel planners to make their vacation a perfect one, and these travel companies depend on Apache Spark for offering various travel packages. TripAdvisor is one such company that uses Apache Spark to compare different travel packages from different providers. It scans through hundreds of websites to find the best and reasonable hotel price, trip package, etc.

Apache Spark: Use Cases

1. Yahoo! has over 1 billion monthly users. It uses Hadoop cluster with more than 40,000 nodes to process data. So, it wanted a lightning-fast computing framework for data processing. Hence, Yahoo! adopted Apache Spark to solve its problem.
2. Although Spark's speed and efficiency is impressive, Yahoo! isn't removing its Hadoop architecture. They need both; Spark will be preferred for real-time streaming and Hadoop will be used for batch processing. The most interesting fact here is that both can be used together through YARN.
3. An American multinational e-commerce corporation, eBay directly connects buyers and sellers. So, a lightning-fast engine is required to handle huge volumes of this real-time streaming data.
4. Apache Spark is mainly used to redefine better customer experience and overall performance at eBay. Apache Spark and Hadoop YARN combine the powerful functionalities of both.

