**Abstract:**

**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.

Spark is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive queries and streaming. Apart from supporting all these workload in a respective system, it reduces the management burden of maintaining separate tools.

**History:**

Spark became a top-level project of the Apache Software Foundation in February 2014, and version 1.0 of Apache Spark was released in May 2014. Spark version 2.0 was released in July 2016.

The technology was initially designed in 2009 by researchers at the University of California, Berkeley as a way to speed up processing jobs in Hadoop systems.

**Pre Requisites for spark:**

· Basic knowledge of python

· scala

· sql

· Linux

· Hadoop

· statistics

**How Apache Spark works:**

Apache Spark can process data from a variety of data repositories, including the Hadoop Distributed File System (HDFS), NoSQL databases and relational data stores, such as Apache Hive. Spark supports in-memory processing to boost the performance of big data analytics applications, but it can also perform conventional disk-based processing when data sets are too large to fit into the available system memory.

The Spark Core engine uses the resilient distributed data set, or RDD, as its basic data type. The RDD is designed in such a way so as to hide much of the computational complexity from users. It aggregates data and partitions it across a server cluster, where it can then be computed and either moved to a different data store or run through an analytic model. The user doesn't have to define where specific files are sent or what computational resources are used to store or retrieve files.

In addition, Spark can handle more than the batch processing applications that MapReduce is limited to running.

**Uses of Spark:**

· Data processing applications

· Batch applications

· SQL

· Machine Learning

· Streaming data processing

· Graph data processing

Nikita Tambe (36075)