

Meeting 4/19/18

Thursday, April 19, 2018 2:14 PM

Attendees: everyone

Hotsos went extremely well

Agenda:

Get familiar with Spark & friends to see how it compares to Oracle

Everything Amazon, spark, data bricks

Hive, parquet

Andy's going to build a VM that we can all use

Notes:

Spark ecosystem: <https://www.dezyre.com/article/apache-spark-ecosystem-and-spark-components/219>



1

Spark Core Component

Responsible for basic I/O functionalities, scheduling and monitoring the jobs on spark clusters, task dispatching, networking with different storage systems, fault recovery and efficient memory management.

Spark SQL Component

Leverage the power of declarative queries and optimized storage by running SQL like queries on Spark data that is present in RDDs and other external sources.

2**3**

Spark Streaming

Allows developers to perform batch processing and streaming of data with ease, in the same application.

MLlib

MLlib eases the deployment and development of scalable machine learning pipelines

4**5**

GraphX

Data scientist can work with graph and non-graph sources to achieve flexibility and resilience in graph construction and transformation.

3

Apache Spark's Cousin Tachyon An in-memory reliable file system

Tachyon is a reliable shared memory that forms an integral part of the Spark ecosystem which helps achieve the desired throughput and performance by avoiding unnecessary replications

achieve the desired throughput and performance by avoiding unnecessary replications.

Cluster Management in Apache Spark

4

1

Standalone
Cluster

Apache
Mesos

2

YARN

3

DeZyre

