Lesson 1

Key Terms Spark Programming, Data and Tabular-Data Processing

Software stack

- Refers to a group of programs, programs work in together or in conjunction, when producing results Also refers to any set of applications that works in a specific and defined order

Example of Software Stack

- LAMP is a software stack that consists of a group of open source components, namely Linux, Apache, MySQL, Perl, PHP or Python

In-memory processing

- Fast when compared to processing data most of the times, from the disk or remotely distributed nodes
- Much less time in accessing the memory compared to the disk or remote data node
- Enables real-time and stream processing

Application-tasks Processing-framework

- Processing with specific data sources such as HDFS as well Hadoop compatible data sources, such as HBase, Cassandra, Ceph, cloud-based Objects Store Service or Amazon S3 and specific programming ways such as Hive, Pig, and other Hadoop ecosystem tools in Java, Python, R and Scala

User Defined Functions (UDFs)

- Refer to custom functions which are not built-in a programming language but user adds them and they can be written in a language, such as Java, Python, Ruby, Jython, JRuby or Scala, and easily embedded functions into scripts in that language

Vectorized UDFs (VUDFs)

- Refer to custom functions using series data-structure, such as one dimensional array or tuples

Grouped Vectorized UDFs (GVUDFs)

- Refer to custom functions written using DataFrame as inputs
- DataFrame examples– A named column consisting of dataset of rows in case of Spark or R

Schema

- Refers to a blueprint for organization or structuring of a database, table or dataset
- Example, database schema tells how the database constructs; the schema defines as set of formulae or just sentences, called integrity constraints, imposed on database

DataFrame in Spark

- Refers to a distributed collection of data that organizes into the named columns,
- A concept similar to a database table in a relational database

Scheme RDD

- Is the name of Spark DataFrame in the earlier versions of Spark

Resilient Distributed Datasets (RDDs)

- A programming abstraction in Spark framework RDDs are also fault tolerant.
- RDD represents fault tolerant dataset storing as collection of Object Stores distributed across many compute nodes for parallel processing.

RDD Partitions

- Storing data in RDD different partitions
- Table has partitions into columns or rows. Similarly, an RDD can also be considered as a table in a database that can hold any type of data.

SerDe

- Refers to Serializer/Deserializer functions (methods)
- Java syntax is SERDE, serde.class.name'
- SerDe use in codes for obtaining records from unstructured data.

Serializer and Deserializer

- Serializer function saves the records, such as columns, rows, tables in serial format
- Deserializer function loads (extracts) the records into format such as columns, rows, tables

Data Pipeline

- Data collected from various data sources that passes through in-between phases (stages) of processing
- The output data of each stage is the input data to the next in the pipeline
- Data Processing in-between uses a chain of function calls in an application or process, such as ETL

ETL

- Process to Extract, Transform and Load (of data)
- .

Shell

- Refers to an environment to write and run the programming scripts
- For example the scripts for query processing similar to SQL

Graph

- Refers to a non-linear data structure with properties attached to each vertex and edge
- Computations perform at each node in a graph structure using path traversals between the vertices (Section 8.2)

Directed Acyclic Graph (DAG)

- Refers to a directed graph with no cyclic traversal
- Here, one set of inputs simultaneously applies at a DAG node input, and after the operations (computations) at the node, only one set of outputs is generated

Nested tables in databases

- Refer to one column tables
- A database storing the rows of a nested table in no particular order
- SQL assigns the rows in consecutive subscripts starting at 1 like an array element [PL/SQL accesses a nested table in no order.]

Parquet

- Refers to a nested hierarchical columnar storage in a group of rows, wherein each row has a number of columns, each column has one chunk, and each chunk has a number of pages
- Page, the minimum processing dataset

Columnar in-memory processing

- Refers to usages of optimized layout in columnar tables (nested tables, Hive, ORC tables)
- Provides the easier data locality using successive memory addresses for the performance for native vectorized optimization

Summary

We learnt meanings of:

- Software Stack
- In-Memory Processing
- Spark
- Columnar Data
- Parquet

End of Lesson 1 on

Key Terms Spark Programming, Data and Tabular-Data Processing