Technologies to Support Distributed Processing

Lesson 10 – Section 2

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Topics

Progression of Distributed Computing:

- Hadoop
- MapReduce
- HBase
- Hive
- Yarn
- Mesos
- Spark

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Overview of Hadoop

It is an ecosystem of programs that supports distributed processing:

- –Hadoop Common
- –Hadoop Distributed File System (HDFS)
- –Hadoop YARN; and
- –Hadoop MapReduce
- –Hive, Pig, and many, many others

Hadoop is not a database system: "append only" write system—no updates without delete and replace

MapReduce Overview

- Provides a parallel programming model for Hadoop (Java-based)
- Moves the computation to the data
- Handled scheduling and fault tolerance across nodes
- Provided status and monitoring of the distributed processes

Advantages and Disadvantages of MapReduce

Advantages

- Simple framework for distributed computation
- All computation must be expressed as a series of steps defined by two simple operations:
  - Mappers – input/sources for data
  - Reducers – writes/sinks for data
- Complete framework for batch jobs expressed as a simple data flow
- Resiliency and fault-tolerance are driven by the nature of the acyclic dataflow graph
- Each map-reduce round is independent of the other, and can be rebuilt if lost
- Communication is handled entirely through the edges in the dataflow graph

Disadvantages

- Each reduce step requires writing back to disk
- For iterative jobs, such as machine learning models, a MapReduce implementation might require multiple reducers
- No global considerations to optimize the dataflow graph

HBase

- Hadoop's database—a distributed, scalable, big data store
- Provides random, real-time read/write access for very large structured tables (modeled on Google's "BigTable") on HDFS
- Linearly and modularly scalable
- Configurable table sharding across servers
- Fault tolerance—failover (within region) and backup support
- Accessible via Java APIs
- Support REST Web services
- Provides Telemetry

Overview of Hive

- Hive is a SQL-like abstraction that allows for interactive queries on top of Hadoop's filesystem
- Has tool support outside of a single supplier—Excel can access Hadoop through Hive
- Hive has a series of Table types it supports: Managed (owned) and External (HDFS)
- New Optimized Row Columnar (ORC) file format is a more efficient storage over HDFS (https://orc.apache.org/)

Hive's Limitations

Disadvantages

- Hive is not fully ANSI SQL – DBs follow this convention—HiveQL does not
- Batch oriented – converts to MapReduce (slow)
- Does "schema on read" which means that they are flexible **AND** fragile—based on Hadoop. If the FS changes it breaks applications built on it.

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON