Midterm Extra Credit

Kieren Gill
ksg7699

What are the components of YARN and what does each component do when a Hadoop MapReduce job is submitted?

VM1 - Master	VM2 - Worker	VM3 - Worker	VM4 - Worker	VMn - Worker
Resource Manager Namenode	Node Manager 1 Datanode 1 ApplicationMaster for Job1 Job1, Task JVM 3	Node Manager 2 Datanode 2 Job 1, Task JVM 1 Job1, Task JVM 2	Node Manager 3 Datanode 3 Job 1, Task JVM 2 Job2, Task JVM 1	Node Manager m Datanode n ApplicationMaster for Job2

What is data locality?

Instead of sending data to where the logic already is -
the logic is sent to where the data already is

Reduces network congestion

Increases overall throughput

Does data locality optimization apply equally to Map tasks and Reduce tasks? Why?

No - it does not apply equally

Mapping:

- **benefit from more optimal placement of logic**
- **data-local is better than rack-local placement**
- **map tasks write data to local disks but read input from HDFS**

Reducing:

- **doesn’t benefit**
- **a single reduce task works on output from multiple map tasks**

What are the uses of Pig, Hive and Impala?

Pig

- language for queries/data manipulation
- uses pig latin
- mapper and reducer code is generated by Pig when compiling MapReduce jobs
- Better for batch processing data
- 1/20 lines of code when compared to MapReduce
- 1/16 development time

What are the uses of Pig, Hive and Impala?

Hive

- built by Facebook
- data warehousing framework (used for reporting and data analysis)
- used for querying and analyzing large datasets in HDFS
- uses HiveQL, similar to SQL
- supports many types of data formats
- uses tables, similar to relational database management systems

What are the uses of Pig, Hive and Impala?

Impala

- **sits on HDFS, doesn’t convert to MapReduce**
- **written in C++**
- **reads Hadoop file formats**
- **can read and write to data files**
- **memory intensive**

What are the differences?

Pig	Hive	Impala
• client side • pig latin • procedural data flow • created by yahoo • Researchers • Programming • .pig extension • no partitioning • does not support schema	• server side • HiveQL • declarative SQL • created by Facebook • Data Analysts • Reports • all extensions • supports partitioning • supports schema • fault tolerant	• created by Cloudera • Impala SQL • restarts if a data node goes down • faster than Hive • can sustain many concurrent users

What does a Mapper do in MapReduce? Explain what the output of a Mapper is.

Mapper

- **input data from users are passed to mapper**
- **format for input is specified**
- **deals with input split**
- **extracts key value pairs**
- **processes input and outputs them in key value pairs**

OUTPUT: Key value pairs sorted by key order

What does a Reducer do in MapReduce? Explain what the output of the Reducer is.

Reducer

- **second stage of processing data**
- **final summation**
- **data aggregation**
- **output will not be sorted**

OUTPUT: Summated/Aggregated Data, can be altered to desired output format

What happens in HDFS when a worker node is corrupt/taken offline? How is data loss prevented? Specifically speak about the HDFS daemons.

Component failures used to be the norm

- had to figure out a way to deal with errors

Fault tolerance

- replicas are created on multiple data nodes throughout the HDFS cluster
- data is accessible from other machines just in case a worker node is corrupt/taken down

What happens in HDFS when a worker node is corrupt/taken offline? How is data loss prevented? Specifically speak about the HDFS daemons.

- **Data Node**

hdfs fsck - command which determines which files have corrupt blocks, gives repair options
command operates only on data

- **Name Node**

./bin/hadoop namenode -recover