BDA - EXP - 3 Word count program using map reduce

Atharva Pawar - 9427 - [Batch - D]

```
cloudera@quickstart:~/workspace
File  Edit  View  Search  Terminal  Help
t.cloudera:8088/proxy/application_1691489656053_0002/
23/08/08 03:26:03 INFO mapreduce.Job: Running job: job_1691489656053_0002
23/08/08 03:26:10 INFO mapreduce.Job: Job job_1691489656053_0002 running in uber
 mode : false
23/08/08 03:26:10 INFO mapreduce.Job:  map 0% reduce 0%
23/08/08 03:26:16 INFO mapreduce.Job:  map 50% reduce 0%
23/08/08 03:26:17 INFO mapreduce.Job:  map 100% reduce 0%
23/08/08 03:26:23 INFO mapreduce.Job:  map 100% reduce 100%
23/08/08 03:26:23 INFO mapreduce.Job: Job job_1691489656053_0002 completed succe
ssfully
23/08/08 03:26:23 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=134
                FILE: Number of bytes written=331676
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=287
                HDFS: Number of bytes written=45
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
```

```
abc.txt
CRCE is my college
I am in CRCE
I am in a college
```

```
[cloudera@quickstart workspace]$ hadoop fs -cat WCOutput/part-00000
CRCE    2
I       2
a       1
am      2
college 2
in      2
is      1
my      1
```

```
[cloudera@quickstart workspace]$ hadoop jar WordCount.jar WCDriver abc2.txt WCOu
tput
23/08/08 03:26:03 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
23/08/08 03:26:03 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
23/08/08 03:26:03 WARN mapreduce.JobSubmitter: Hadoop command-line option parsin
g not performed. Implement the Tool interface and execute your application with
ToolRunner to remedy this.
23/08/08 03:26:03 INFO mapred.FileInputFormat: Total input paths to process : 1
23/08/08 03:26:03 INFO mapreduce.JobSubmitter: number of splits:2
23/08/08 03:26:03 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
91489656053_0002
23/08/08 03:26:03 INFO impl.YarnClientImpl: Submitted application application_16
91489656053 0002
```

hadoop fs -put WCFile.txt WCFile.txt
hadoop jar WordCount.jar WCDriver WCFile.txt WCOutput
hadoop fs -cat WCOutput/part-00000

Atharva Prashant Powar (9427) — [Batch-D]

BDA Exp 3

Q1. Distinguish the Hadoop Ecosystem?

* Hadoop Ecosystem:

A. Hadoop Core Components:
HDFS (Hadoop Distributed file System):
Storage system that stores data accross multiple machines.
MapReduce:
Programming model & processing framework for distributed data processing.

B. Data Storage & Processing:
HBase:
Distributed NoSQL database for real-time read/write access.
Apache Hive: Data warehousing tool for querying & analyzing large datasets using SQL-like queries.
Pig:
High level scripting platform for creating MapReduce programs without writing Java code.
Apache Spark:
In-memory data processing framework for faster analytics & machine learning.

C. Data Ingestion & Integration:
Flume: Collects, aggregates, & moves large amounts of log data from different sources to Hadoop.
Sqoop: Imports data from relational databases into Hadoop & exports data from Hadoop to database.

(2)

D. Data Processing & Analytics:

Apache Kafka:
Publish-subscribe messaging sys. for real-time data streaming & processing.

Apache Storm:
Real time stream processing framework for handling high-velocity data.

Flink:
Distributed stream processing framework for real-time analytics.

Q2) Divide & Conquer in Hadoop Cluster:

=> 1. Problem Overview:
Imagine we have a large dataset that needs processing such as analyzing customers reviews for sentiment analysis.
The dataset is too big to be processed on a single machine due to its size

2. Divide Phase:
The dataset is divided into smaller chunks or partitions.
Each partition contains a portion of the overall data.

3. Distribute Phase:
The partitions are distributed across multiple machines in the Hadoop cluster.
This distribution takes advantage of the cluster's parallel processing capabilities.

4. Conquer Phase:
Each machine process its assigned partitions independently.
For sentiment analysis, each machine would analyze the sentiment of reviews within its partition.

5. Merge Phase:
Once individual machines finish processing, their results are collected & combined.
In our scenario, the sentiment analysis results from all partitions are merged.

6. Final Result:
The final result of sentiment analysis is obtained by combining the sentiment scores from all partitions. This final result provides insights into the overall sentiment of the customer reviews.

7. Advantages of Divide & Conquer in Hadoop:
Enables efficient processing of large datasets by distributing the workload across the cluster.
Utilizes parallel processing, reducing the time required for analysis.
Can handle complex tasks that would be impractical to run on a single machine.