



Introduction to Hadoop Programming

Bryon Gill, Pittsburgh Supercomputing Center

What We Will Discuss



- Hadoop Architecture Overview
- Practical Examples
 - “Classic” Map-Reduce
 - Hadoop Streaming
- Spark, Hbase and Other Applications

Hadoop Overview



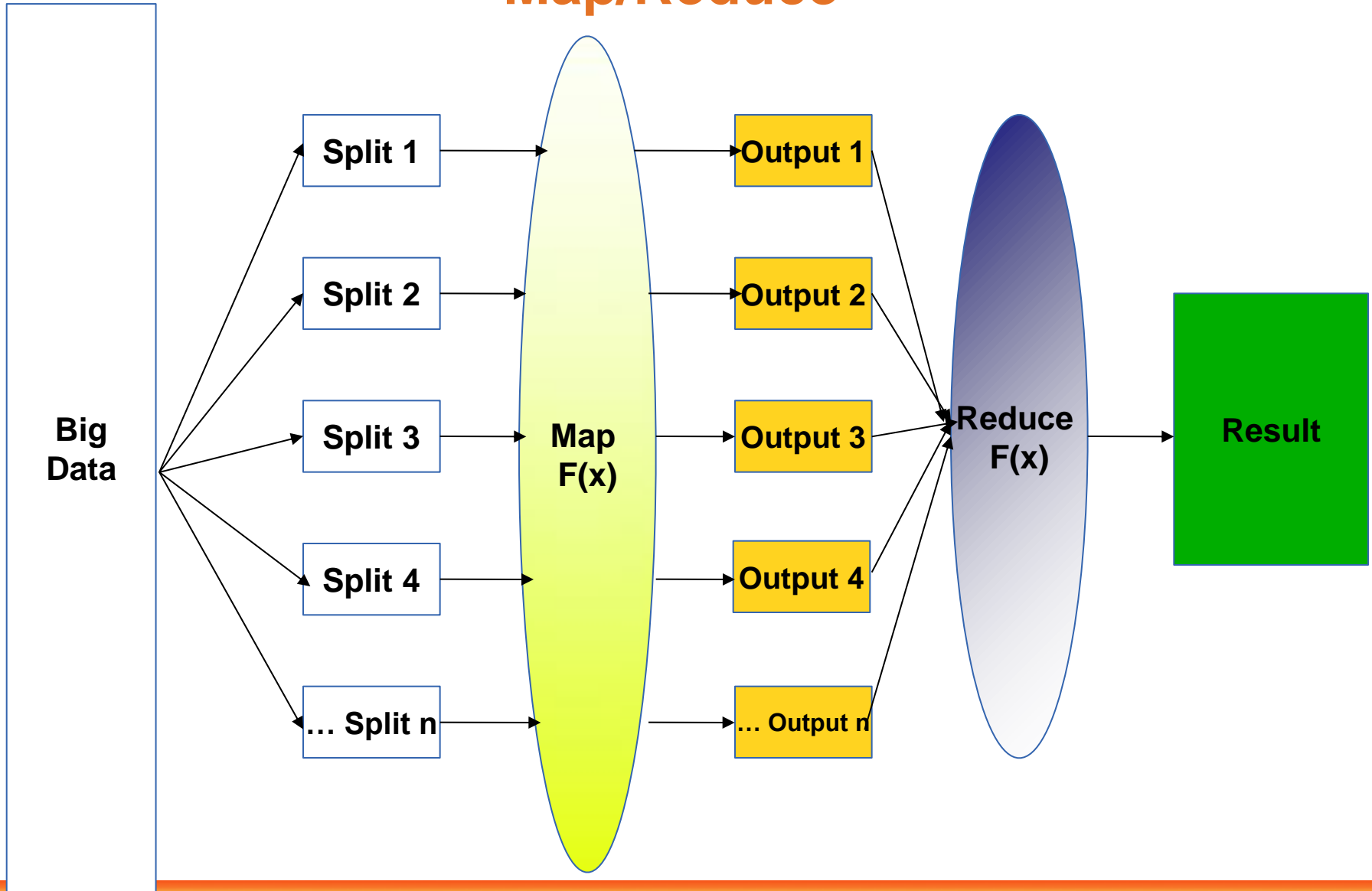
- Framework for Big Data
- Map/Reduce
 - (<http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf>)
- Platform for Big Data Applications

Map/Reduce



- Apply a Function to all the Data
- Harvest, Sort, and Process the Output
- (cat | grep | wc -l)

Map/Reduce



HDFS



- Distributed FS Layer
- WORM fs
 - Optimized for Streaming Throughput
- Exports
- Replication
- Process data in place

HDFS Invocations: Getting Data In and Out



- `hdfs dfs -ls`
- `hdfs dfs -put`
- `hdfs dfs -get`
- `hdfs dfs -rm`
- `hdfs dfs -mkdir`
- `hdfs dfs -rmdir`

Exercise: HDFS



```
hdfs dfs -put /home/training/BigData/* /user/$USER
```


Writing Hadoop Programs



- Wordcount Example: `Wordcount.java`
 - Map Class
 - Reduce Class

Exercise: Compiling



```
cp /home/training/hadoop/WordCount.java ./
hadoop com.sun.tools.javac.Main WordCount.java
```

Exercise: Packaging



```
jar cf wc.jar WordCount*.class
```

Exercise: Submitting



- `hadoop \`
`jar wc.jar \`
`WordCount \`
`/datasets/compleat.txt \`
`output \`
`-D mapred.reduce.tasks=2`

Configuring your Job Submission



- Mappers and Reducers
- Java options
- Other parameters

Monitoring



- Web Interface Ports:
 - `r650.pvt.bridges.psc.edu:8088` – Yarn Resource Manager (Track Jobs)
 - `r650.pvt.bridges.psc.edu:50070` – HDFS (Namenode)
 - `r650.pvt.bridges.psc.edu:19888` – Job History Server Interface

Troubleshooting



- Read the stack trace
- Check the logs!
- Check system levels (disk, memory etc)
- Change job options memory etc.

Hadoop Streaming



- Alternate method for programming MR jobs
- Write Map/Reduce Jobs in any language
- Map and Reduce each read from stdin
- Text class default for input/output (\t or whole line)
- Excellent for Fast Prototyping

Hadoop Streaming: Bash Example



- Bash wc and cat
- `hadoop jar \`
`$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming*.jar \`
`-input /datasets/plays/ \`
`-output streaming-out \`
`-mapper '/bin/cat' \`
`-reducer '/usr/bin/wc -l'`

Hadoop Streaming Python Example



- Wordcount in python
- ```
hadoop jar
$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming*.jar \
-file ~training/hadoop/mapper.py \
-mapper mapper.py \
-file ~training/hadoop/reducer.py \
-reducer reducer.py \
-input /datasets/plays/ \
-output pyout
```

# Questions?



- Thanks!

# References and Useful Links



- HDFS shell commands:  
<http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html>
- Apache Hadoop Official Releases:  
<https://hadoop.apache.org/releases.html>