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

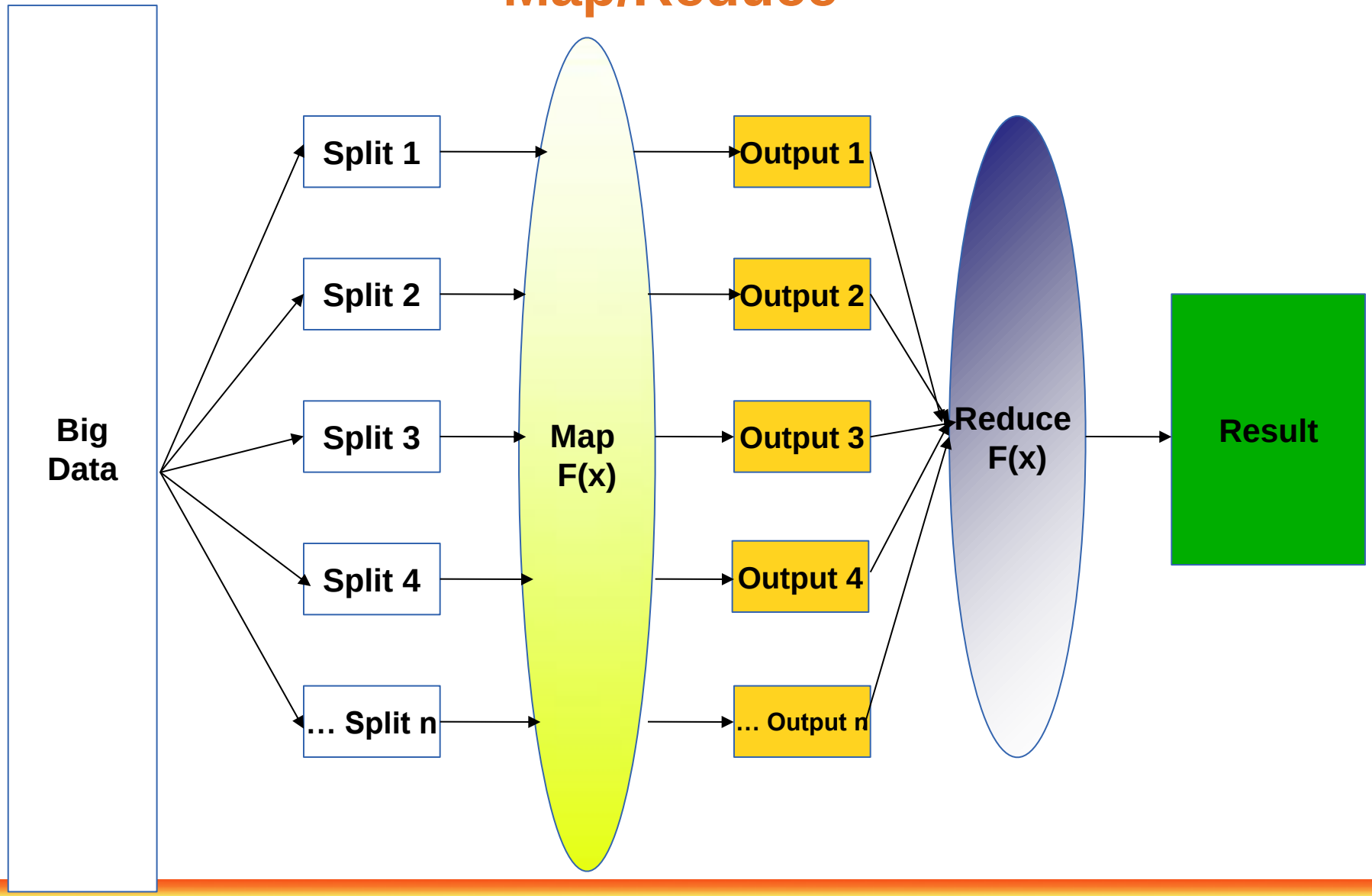
- Framework for Big Data
- Map/Reduce
- (http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf)
- Platform for Big Data Applications

Map/Reduce

PSC
PITTSBURGH SUPERCOMPUTING CENTER

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

Map/Reduce



PSC PITTSBURGH SUPERCOMPUTING CENTER

HDFS

PSC
PITTSBURGH SUPERCOMPUTING CENTER

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

HDFS Invocations: Getting Data In and Out

PSC
PITTSBURGH SUPERCOMPUTING CENTER

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

Writing Hadoop Programs

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

Compiling

PSC
PITTSBURGH SUPERCOMPUTING CENTER

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

Packaging

jar cf wc.jar WordCount*.class

Submitting

PSC
PITTSBURGH SUPERCOMPUTING CENTER

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

Configuring your Job Submission

PSC
PITTSBURGH SUPERCOMPUTING CENTER

- Mappers and Reducers
- Java options
- Other parameters

Monitoring

PSC
PITTSBURGH SUPERCOMPUTING CENTER

- Web Interface Ports (requires proxy on Bridges):
  - r741.pvt.bridges.psc.edu:8088 – Yarn Resource Manager (Track Jobs)
  - r741.pvt.bridges.psc.edu:50070 – HDFS (Namenode)
  - r741.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

PSC
PITTSBURGH SUPERCOMPUTING CENTER

- 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

PSC
PITTSBURGH SUPERCOMPUTING CENTER

- 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

Connecting to the Training Cluster


```
# connect to the login node
ssh bridges.psc.xsede.org
# connect from there to the cluster
ssh r741
# copy the example scripts
cp ~training/hadoop/* .
# run hadoop commands
hdfs dfs –ls /datasets/
```

Connecting to the Training Cluster

PSC
PITTSBURGH SUPERCOMPUTING CENTER

Using the Web Proxy:

Bridges compute nodes can't be reached directly, a proxy must be used following these directions:

https://www.psc.edu/bridges/user-guide/hadoop-and-spark/proxy-set-up-guide

The password for the proxy will be announced during the lecture.