Software Requirements

- Scala
- Java
- Spark
- Maven *
- Sbt *
- Python *
- R *
- Hadoop *

```
* Depends on user preference.
  Java, Python, R API availability lags behind Scala
```

API

- RDDs
- Resilient Distributed Datasets
  - Main abstraction in Spark
  - Distributed, fault-tolerant collection of partitioned records on multiple nodes
- Transformations
  - Create new RDDs from existing RDDs
- Actions
  - Kicks off execution

Spark Application

- Spark driver => coordinates spark processes processes
- Spark processes => run independently on cluster
- A Spark application is combination of driver and processes
- Note that actions like ‘collect()’ can bring back too much data to the driver

API Levels - RDD

- Low level, compile-time type-safe
- Expresses how instead of what
- Functional programming constructs
- No query optimization done for us
- Can build inefficient RDD transformation chains (ReduceByKey – counting hole dataset, before a filtering) rdd.reduceByKey().filter {etc}

API Levels - DataFrame

- Relational database feel
- Domain specific language to manipulate data
- Expresses what instead of how
- Schema is imposed (columns with names and types)
- Queries are optimized by Spark engine
- No type safety

API Levels - Dataset

- New abstraction, higher level API
- Merges with DataFrames API
- Adds compile-time errors
- Errors are raised before jobs kid off
- More Info :
  - https://www.youtube.com/watch?v=pZQsDloGB4w

Spark Application

- Spark driver => coordinates spark processes processes
- Spark processes => run independently on cluster
- A Spark application is combination of driver and processes
- Note that actions like ‘collect()’ can bring back too much data to the driver

Spark Shell

- Modified Scala shell
- spark-shell --master local[*]
  - Master URL for distributed cluster OR local[N] to run locally with N threads
  - Ex:
    - spark-shell -- master spark://spark-master:7077
    - spark-shell -- master local[3]
  - http://localhost:4040 <- worker
  - http://spark-master:8080 <-cluster & workers

Spark Shell

Start Spark locally with 10 threads

```
richardr@spark-master:~/dev/git/spark-helloworld
File  Edit  View  Search  Terminal  Help
[richardr@spark-master spark-helloworld]$ spark-shell --master local[10]
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
17/02/06 10:08:14 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin
-java classes where applicable
17/02/06 10:08:21 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Spark context Web UI available at http://132.246.236.99:4040
Spark context available as 'sc' (master = local[10], app id = local-1486390095485).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.1.0
      /_/

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_112)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

Spark WebUI

(spark-shell with 10 threads)

Spark shell - Executors - Mozilla Firefox

Spark shell - Executors

localhost:4040/executors/

Most Visited | Internal_Dev | Fedora Documentation | Fedora Project | Red Hat | Free Content | From Google Chrome

Spark 2.1.0 — Jobs | Stages | Storage | Environment | Executors | SQL — Spark shell application UI

Executors

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write
Active(1)	0	0.0 B / 384.1 MB	0.0 B	10	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B
Total(1)	0	0.0 B / 384.1 MB	0.0 B	10	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B

Executors

Show 20 entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	T
driver	132.246.236.99:45887	Active	0	0.0 B / 384.1 MB	0.0 B	10	0	0	0	0	0

Showing 1 to 1 of 1 entries Previous 1 Next

http://spark.apache.org/docs/latest/programming-guide.html#using-the-shell

Example – Scala, RDD, Spark Shell

```
//Read in a text file :
val lines = sc.textFile("/home/richardr/Documents/data/inputdata.txt")

//Actions : Show first element of the lines RDD and print all lines
lines.first()
lines.foreach(println)

//Split each lines on blan space
val words = lines.flatMap(ln => ln.split(" "))

//Actions : Show first element of words RDD and print all words
words.first()
words.foreach(println) //prints array of strings

//Create a pairs RDD, first element of pair is the word, second is
//the number 1
val pairs = words.map(w => (w, 1))
pairs.foreach(println)

//Count words (grouped by word). Key is word, count is
//second element of pair
val counts = pairs.reduceByKey(_ + _)

//Create RDD of string representations for pair RDD
val strPairs = counts.map(tuple => tuple.productIterator.mkString(","))

//Print each word and count
strPairs.foreach(println)
```

* For Java Version, See : http://spark.apache.org/examples.html

Spark Cluster

- Local[*] <- Great for development
- Spark://[a spark master]:8080 <- larger jobs
- Starting a cluster
  - conf/slaves
  - conf/spark-env.sh
  - sbin/start-all.sh
  - sbin/start-slave.sh spark://[a spark master]:7077

Spark Cluster

```
[richardr@spark-master sbin]$ more ../conf/slaves
#
# Licensed to the Apache Software Foundation (ASF) under one or more
# contributor license agreements.  See the NOTICE file distributed with
# this work for additional information regarding copyright ownership.
# The ASF licenses this file to You under the Apache License, Version 2.0
# (the "License"); you may not use this file except in compliance with
# the License.  You may obtain a copy of the License at
#
#    http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.
#

# A Spark Worker will be started on each of the machines listed below
spark-master
spark-slave-1
spark-slave-2
[richardr@spark-master sbin]$ sh start-all.sh
starting org.apache.spark.deploy.master.Master, logging to /home/richardr/apps/spar
ark-richardr-org.apache.spark.deploy.master.Master-1-spark-master.out
spark-slave-1: starting org.apache.spark.deploy.worker.Worker, logging to /home/rich
doop2.7/logs/spark-richardr-org.apache.spark.deploy.worker.Worker-1-vmbox.out
spark-master: starting org.apache.spark.deploy.worker.Worker, logging to /home/richa
oop2.7/logs/spark-richardr-org.apache.spark.deploy.worker.Worker-1-spark-master.out
spark-slave-2: starting org.apache.spark.deploy.worker.Worker, logging to /home/rich
doop2.7/logs/spark-richardr-org.apache.spark.deploy.worker.Worker-1-NRC-005613.out
```

Slave Configuration

Slaves need Spark software

Spark Master at spark://spark-master:7077 - Mozilla Firefox

Spark Master at spark://s...

localhost:8080

Search

Most Visited | Internal_Dev | Fedora Documentation | Fedora Project | Red Hat | Free Content | From Google Chrome

Spark 2.1.0

Spark Master at spark://spark-master:7077

URL: spark://spark-master:7077
REST URL: spark://spark-master:6066 *(cluster mode)*
Alive Workers: 3
Cores in use: 36 Total, 0 Used
Memory in use: 36.0 GB Total, 0.0 B Used
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers

Worker Id	Address	State	Cores	Memory
worker-20170206104613-132.246.236.113-36007	132.246.236.113:36007	ALIVE	8 (0 Used)	8.0 GB (0.0 B Used)
worker-20170206104614-132.246.236.99-41467	132.246.236.99:41467	ALIVE	20 (0 Used)	20.0 GB (0.0 B Used)
worker-20170206104619-132.246.236.109-40277	132.246.236.109:40277	ALIVE	8 (0 Used)	8.0 GB (0.0 B Used)

Running Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration

Completed Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration

3 Workers with various configurations

Submitting a Job

- For self-contained applications
- deploy-mode -- local <- driver runs locally (default)
- deploy-mode -- cluster <- driver runs on cluster (fire and forget)

spark-submit

```
#!/bin/bash
spark-submit \
    --class ca.redsofa.jobs.HelloWorld \
    --master local[*] \
    ./target/spark-helloworld-1.0-SNAPSHOT.jar
```

Submit locally
With :
threads = # cores
Driver memory 1G default

```
#!/bin/bash
spark-submit --class ca.nrc.jobs.App \
--master spark://spark-master:7077 \
--driver-memory 2G \
./target/batch-convert-json-to-parquet-0.1-SNAPSHOT.jar \
file:///home/richardr/data/qcr_data/Olympics_Data/Olympics_corenlp_annotated_excerpt.json \
file:///home/richardr/data/qcr_data/Olympics_Data/Olympics_corenlp_annotated_excerpt_field_subset.json
```

Submit to cluster
With :
workers => cluster managed
Driver memory 2G

```
#!/bin/bash
spark-submit --class ca.nrc.jobs.App \
--master spark://spark-master:7077 \
--deploy-mode cluster \
--driver-memory 2G \
--conf 'spark.executor.memory=8g' \
hdfs://hadoop-master:9000/home/richardr/jars/batch-corenlp-annotation-0.1-SNAPSHOT.jar \
hdfs://hadoop-master:9000/home/richardr/data/Ottawa_Shooting/ottawa_shooting_tweets.json \
hdfs://hadoop-master:9000/home/richardr/data//Ottawa_Shooting/OS_cluster_batch_corenlp_annotated.json
```

Submit to cluster
With :
workers => cluster managed
Driver process runs on cluster
Nodes need to see everything

Self-Contained Applications Java, Dataset, spark-submit

```java
//1 - Start the Spark session
SparkSession spark = SparkSession
        .builder()
        .appName("Simple Batch Job")
        .config("spark.driver.memory", "2g")
        .enableHiveSupport()
        .getOrCreate();

//2 - Read in the text file
Dataset<String> inputDataDs = spark.read().text(INPUT_FILE).as(Encoders.STRING());

//3 - Create words data set. Take each line in the inputDataDs and create one row
// for each word in the text file.

// Source : https://gist.github.com/lucianogiuseppe/063aff936f548fdd0faad6ef004a43e7
Dataset<String> words = inputDataDs.flatMap(s -> {
                            return  Arrays.asList(s.toLowerCase().split(" ")).iterator();
                        }, Encoders.STRING())
                        .filter(s -> !s.isEmpty());

words.printSchema();


//4 - Create a temporary table so we can use SQL queries
words.createOrReplaceTempView("words");

//5 - Write and execute query
String sql = "SELECT " +
                "value as word, " +
                "COUNT(value) as word_count " +
             "FROM " +
               "words " +
             "GROUP BY " +
               "value " +
             "ORDER BY " +
                "value " +
             "ASC " ;

Dataset<Row> wordCount = spark.sql(sql);
```

Reading a text file and splitting lines on Spaces

Note :

If reading JSON, the schema is automatically inferred

```
[richardr@spark-master spark-helloworld]$ sh submit_job.sh
Stating Job...
17/02/06 10:22:53 WARN NativeCodeLoader: Unable to load native-hadoop library
-java classes where applicable
17/02/06 10:22:59 WARN ObjectStore: Failed to get database global_temp, retu
root
 |-- value: string (nullable = true)

+-----+----------+
| word|word_count|
+-----+----------+
|    a|         1|
|again|         2|
|hello|         2|
|   is|         1|
|  one|         2|
| test|         1|
| this|         1|
|three|         2|
|  two|         2|
|world|         2|
+-----+----------+
```

* Source here : https://github.com/redsofa/spark-helloworld

User Defined Functions

```java
public static void registerStringLengthUdf(SparkSession spark){
    spark.udf().register("stringLengthUdf", new UDF1<String, Long>() {
      @Override
      public Long call(String str) {
          if(str != null && !str.isEmpty()){
            return new Long(str.length());
          }else{
            return 0L;
          }
        }
    }, DataTypes.LongType);
}
```

Any Java code

```java
    //7 - Register user defined function (UDF)
    registerStringLengthUdf(spark);
    wordCount.createOrReplaceTempView("word_counts");

    //8 - Call UDF from a SQL statement.
    //    Note : This UDF could be called from R, Python, Java and Scala code
    sql = "SELECT *, stringLengthUdf(word) as my_UDF_str_length FROM word_counts";

    Dataset <Row> wordCountWithLengths = spark.sql(sql);
    wordCountWithLengths.show(10);


    //9 - Call UDF (again) but now by using the withColumn method
    Dataset <Row> withColumnDs = wordCount
                                    .withColumn("my_UDF_str_len",
                                      callUDF("stringLengthUdf", wordCount.col("word"))
                                    );
```

SparkR

spark_convert.r

```r
sparkInstallDir <- "/home/richardr/apps/spark-2.1.0-bin-hadoop2.7"
inputFile <- "/home/richardr/data/example/ppl.json"
applicationName <- "Convert_ETL" #Application name to register with cluster manager
rdataFile <- "/home/richardr/data/example/ppl.RDATA"

Sys.setenv(SPARK_HOME = sparkInstallDir)
library(SparkR, lib.loc = c(file.path(Sys.getenv("SPARK_HOME"), "R", "lib")))

sparkR.session(master = "local[10]", appName = applicationName,
               sparkHome = Sys.getenv("SPARK_HOME"), sparkConfig = list(spark.driver.memory="2g"),
               sparkJars = "", sparkPackages = "", enableHiveSupport = TRUE)

# Load the input JSON file
inputData <- read.df(inputFile, "json")

printSchema(inputData)
```

ppl.json

```
{"firstName": "Joey", "lastName": "Jollymore", "sex": "M", "age": 12}
{"firstName": "Walter", "lastName": "Hasthag", "sex": "M", "age": 21}
{"firstName": "Willy", "lastName": "Waller", "sex": "M", "age": 11}
{"firstName": "Anna", "lastName": "Walsh", "sex": "F", "age": 33}
{"firstName": "Wonder", "lastName": "Woman", "sex": "F", "age": 34}
```

JSON Schema inference

Need to tell R where Spark is installed

Console ~/dev/git/dac-ast-2016-drayage/src/shiny-app/POM-UI/

```
    cov, filter, lag, na.omit, predict, sd, var, window

The following objects are masked from 'package:base':

    as.data.frame, colnames, colnames<-, drop, endsWith, intersect, rank, rbind, sample, startsWith, subset, summary,
    transform, union

>
> sparkR.session(master = "local[10]", appName = applicationName,
+                sparkHome = Sys.getenv("SPARK_HOME"), sparkConfig = list(spark.driver.memory="2g"),
+                kages = "", enableHiveSupport = TRUE)
                richardr/apps/spark-2.1.0-bin-hadoop2.7
                /home/richardr/apps/spark-2.1.0-bin-hadoop2.7/bin/spark-submit   --driver-memory
                d_port43d4730d3ad1
                apache/spark/log4j-defaults.properties

                (newLevel). For SparkR, use setLogLevel(newLevel).
                Unable to load native-hadoop library for your platform... using builtin-java cla

                on information not found in metastore. hive.metastore.schema.verification is not
                .2.0
17/02/06 11:35:22 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException
17/02/06 11:35:26 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Java ref type org.apache.spark.sql.SparkSession id 1
>
> # Load the input JSON file
> inputData <- read.df(inputFile, "json")
>
> printSchema(inputData)
root
 |-- age: long (nullable = true)
 |-- firstName: string (nullable = true)
 |-- lastName: string (nullable = true)
 |-- sex: string (nullable = true)
>
```

Start Spark locally with 10 threads

Closure Serialization

- "Task not serializable" exceptions
- Need to be mindful about executing code across a cluster
- Closures are serialized and sent to each cluster executor
- See Understanding Closures

Spark on Dorval Cluster

!!! EXPERIMENTAL – EARLY RESULTS !!!

Spark on Dorval Cluster - jobfile

gpsc-in.science.gc.ca - PuTTY

```
#!/bin/bash
#
#
#$ -N spark_test_job
#

#$ -j y
#$ -pe dev 10
#$ -l res_cpus=10
#$ -l res_mem=16000
#$ -l h_rt=86400
#$ -l res_image=nrc_all_default_centos-6.7-amd64_latest

set -x

echo "The master... "
echo $(hostname)

echo "The slaves... "
cat $GECOSHEP_PNET_HOSTFILE | grep -v $(hostname)

echo "Cleaning up logs and slaves fle ..."
rm -f  $HOME/u/apps/spark-2.1.0-bin-hadoop2.7/logs/*
rm -f $HOME/u/apps/spark-2.1.0-bin-hadoop2.7/conf/slaves

echo "Writing the slaves file ..."
for HOST in $(cat $GECOSHEP_PNET_HOSTFILE | grep -v $(hostname));
do
   echo $HOST >> $HOME/u/apps/spark-2.1.0-bin-hadoop2.7/conf/slaves
done

echo "Starting Spark Cluster ..."
ssh $(hostname) 'sh $HOME/u/apps/spark-2.1.0-bin-hadoop2.7/sbin/start-all.sh'

echo "Submitting Job ..."
ssh $(hostname) 'source $HOME/.profile.rir002; cd $HOME/u/spark_jobs/helloworld; spark-submit --class ca.nrc.App --master=spark://$(hostname):7077 $HOME/u/spark_jobs/helloworld/target/helloworld-1.
0-SNAPSHOT.jar'

echo "Cleaning up slaves file ..."
rm -f $HOME/u/apps/spark-2.1.0-bin-hadoop2.7/conf/slaves
~
~
~
~
~
~
~
~
~
~
"_10x10x10x16GB.job" 39L, 1074C                                                    8,1          All
```

!!! EXPERIMENTAL – EARLY RESULTS !!!

Start

gpsc-in.science.g... 85% Uploading - r... 11:08 AM 2/6/2017

Spark on Dorval Cluster – submit

gpsc-in.science.gc.ca - PuTTY

```
bash-4.1$ jobsub .jobfiles/spark_10x10x10x16GB.job
warning: "-N" option has already been set, overriding previous setting
Your job 2925247 ("spark_test_job") has been submitted
bash-4.1$
```

!!! EXPERIMENTAL – EARLY RESULTS !!!

Start

gpsc-in.science.g... 86% Uploading - r...

11:10 AM
2/6/2017

Spark on Dorval Cluster - log

gpsc-in.science.gc.ca - PuTTY

```
bash-4.1$ ls
Desktop  spark_test_job.o2889870  spark_test_job.o2890227  spark_test_job.o2890345  spark_test_job.o2925231  spark_test_job.o2925247  u
bash-4.1$
```

!!! EXPERIMENTAL – EARLY RESULTS !!!

Start

gpsc-in.science.g...

88% Uploading - r...

11:14 AM
2/6/2017

Spark on Dorval Cluster - log

gpsc-in.science.gc.ca - PuTTY

```
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4150.0 in stage 0.0 (TID 4150) in 3191 ms on 10.107.191.13 (executor 3) (4143/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4431.0 in stage 0.0 (TID 4431, 10.107.135.205, executor 5, partition 4431, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4165.0 in stage 0.0 (TID 4165) in 2977 ms on 10.107.135.205 (executor 5) (4144/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4432.0 in stage 0.0 (TID 4432, 10.107.191.13, executor 3, partition 4432, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4166.0 in stage 0.0 (TID 4166) in 2953 ms on 10.107.191.13 (executor 3) (4145/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4433.0 in stage 0.0 (TID 4433, 10.107.156.139, executor 7, partition
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4143.0 in stage 0.0 (TID 4143) in 3284 ms on 10.107.156.139 (executor
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4434.0 in stage 0.0 (TID 4434, 10.107.129.9, executor 1, partition 44
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4135.0 in stage 0.0 (TID 4135) in 3384 ms on 10.107.129.9 (executor 1) (4147/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4435.0 in stage 0.0 (TID 4435, 10.107.180.65, executor 0, partition 4435, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4161.0 in stage 0.0 (TID 4161) in 3112 ms on 10.107.180.65 (executor 0) (4148/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4436.0 in stage 0.0 (TID 4436, 10.107.180.65, executor 0, partition 4436, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4188.0 in stage 0.0 (TID 4188) in 2829 ms on 10.107.180.65 (executor 0) (4149/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4437.0 in stage 0.0 (TID 4437, 10.107.156.139, executor 7, partition 4437, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4117.0 in stage 0.0 (TID 4117) in 3677 ms on 10.107.156.139 (executor 7) (4150/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4438.0 in stage 0.0 (TID 4438, 10.107.195.128, executor 6, partition 4438, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4140.0 in stage 0.0 (TID 4140) in 3372 ms on 10.107.195.128 (executor 6) (4151/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4439.0 in stage 0.0 (TID 4439, 10.107.129.9, executor 1, partition 4439, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4128.0 in stage 0.0 (TID 4128) in 3559 ms on 10.107.129.9 (executor 1) (4152/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4440.0 in stage 0.0 (TID 4440, 10.107.191.13, executor 3, partition 4440, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4151.0 in stage 0.0 (TID 4151) in 3348 ms on 10.107.191.13 (executor 3) (4153/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4441.0 in stage 0.0 (TID 4441, 10.107.195.128, executor 6, partition 4441, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4120.0 in stage 0.0 (TID 4120) in 3727 ms on 10.107.195.128 (executor 6) (4154/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4442.0 in stage 0.0 (TID 4442, 10.107.156.139, executor 7, partition 4442, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4137.0 in stage 0.0 (TID 4137) in 3477 ms on 10.107.156.139 (executor 7) (4155/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4443.0 in stage 0.0 (TID 4443, 10.107.135.205, executor 5, partition 4443, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4154.0 in stage 0.0 (TID 4154) in 3302 ms on 10.107.135.205 (executor 5) (4156/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4444.0 in stage 0.0 (TID 4444, 10.107.135.205, executor 5, partition 4444, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4157.0 in stage 0.0 (TID 4157) in 3304 ms on 10.107.135.205 (executor 5) (4157/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4445.0 in stage 0.0 (TID 4445, 10.107.191.13, executor 3, partition 4445, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4144.0 in stage 0.0 (TID 4144) in 3457 ms on 10.107.191.13 (executor 3) (4158/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4446.0 in stage 0.0 (TID 4446, 10.107.156.139, executor 7, partition 4446, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4177.0 in stage 0.0 (TID 4177) in 3084 ms on 10.107.156.139 (executor 7) (4159/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4447.0 in stage 0.0 (TID 4447, 10.107.180.65, executor 0, partition 4447, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4214.0 in stage 0.0 (TID 4214) in 2719 ms on 10.107.180.65 (executor 0) (4160/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4448.0 in stage 0.0 (TID 4448, 10.107.129.9, executor 1, partition 4448, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4175.0 in stage 0.0 (TID 4175) in 3100 ms on 10.107.129.9 (executor 1) (4161/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4449.0 in stage 0.0 (TID 4449, 10.107.181.78, executor 8, partition 4449, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4146.0 in stage 0.0 (TID 4146) in 3453 ms on 10.107.181.78 (executor 8) (4162/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4450.0 in stage 0.0 (TID 4450, 10.107.197.72, executor 2, partition 4450, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4149.0 in stage 0.0 (TID 4149) in 3450 ms on 10.107.197.72 (executor 2) (4163/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4451.0 in stage 0.0 (TID 4451, 10.107.156.139, executor 7, partition 4451, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4158.0 in stage 0.0 (TID 4158) in 3352 ms on 10.107.156.139 (executor 7) (4164/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4452.0 in stage 0.0 (TID 4452, 10.107.136.206, executor 4, partition 4452, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4145.0 in stage 0.0 (TID 4145) in 3491 ms on 10.107.136.206 (executor 4) (4165/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4453.0 in stage 0.0 (TID 4453, 10.107.136.206, executor 4, partition 4453, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4129.0 in stage 0.0 (TID 4129) in 3684 ms on 10.107.136.206 (executor 4) (4166/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4454.0 in stage 0.0 (TID 4454, 10.107.180.65, executor 0, partition 4454, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4206.0 in stage 0.0 (TID 4206) in 2848 ms on 10.107.180.65 (executor 0) (4167/7074)
17/02/06 15:13:58 INFO TaskSetManager: Starting task 4455.0 in stage 0.0 (TID 4455, 10.107.129.9, executor 1, partition 4455, PROCESS_LOCAL, 6053 bytes)
17/02/06 15:13:58 INFO TaskSetManager: Finished task 4187.0 in stage 0.0 (TID 4187) in 3070 ms on 10.107.129.9 (executor 1) (4168/7074)
```

!!! EXPERIMENTAL – EARLY RESULTS !!!

gpsc-in.science.g... | 88% Uploading - r... | Untitled - Paint | 11:14 AM 2/6/2017

Spark on Dorval Cluster - webui

gpsc-in.science.gc.ca - PuTTY

```
Desktop  spark_test_job.o2889870  spark_test_job.o2890227  spark_test_job.o2890345  spark_test_job.o2925231  spark_test_job.o2925247  u
bash-4.1$ ssjh -j 2925321 -- firefox
bash: ssjh: command not found
bash-4.1$ sshj -j 2925321 -- firefox
Warning: Permanently added 'ib7-bc2ll42-be02p14.science.gc.ca,10.107.133.13' (RSA) to the list of known hosts.
```

Spark Master at spark://ib7-bc2ll42-be02p14.science.gc.ca:7077 - Mozilla Firefox

Spark Master at spark://ib7-...

localhost:8080

Search

Spark 2.1.0 Spark Master at spark://ib7-bc2ll42-be02p14.science.gc.ca:7077

URL: spark://ib7-bc2ll42-be02p14.science.gc.ca:7077
REST URL: spark://ib7-bc2ll42-be02p14.science.gc.ca:6066 *(cluster mode)*
Alive Workers: 9
Cores in use: 288 Total, 288 Used
Memory in use: 1124.4 GB Total, 18.0 GB Used
Applications: 1 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

!!! EXPERIMENTAL – EARLY RESULTS !!!

Workers

Worker Id	Address	State	Cores	Memory
worker-20170206152246-10.107.129.9-35859	10.107.129.9:35859	ALIVE	32 (32 Used)	124.9 GB (2.0 GB Used)
worker-20170206152247-10.107.136.206-48228	10.107.136.206:48228	ALIVE	32 (32 Used)	124.9 GB (2.0 GB Used)
worker-20170206152247-10.107.156.139-58970	10.107.156.139:58970	ALIVE	32 (32 Used)	124.9 GB (2.0 GB Used)
worker-20170206152247-10.107.180.65-45679	10.107.180.65:45679	ALIVE	32 (32 Used)	124.9 GB (2.0 GB Used)
worker-20170206152247-10.107.181.78-33876	10.107.181.78:33876	ALIVE	32 (32 Used)	124.9 GB (2.0 GB Used)
worker-20170206152247-10.107.191.13-54298	10.107.191.13:54298	ALIVE	32 (32 Used)	124.9 GB (2.0 GB Used)
worker-20170206152247-10.107.197.72-53480	10.107.197.72:53480	ALIVE	32 (32 Used)	124.9 GB (2.0 GB Used)
worker-20170206152248-10.107.131.64-34265	10.107.131.64:34265	ALIVE	32 (32 Used)	124.9 GB (2.0 GB Used)
worker-20170206152248-10.107.194.68-44483	10.107.194.68:44483	ALIVE	32 (32 Used)	124.9 GB (2.0 GB Used)

Running Applications