Spark and Parquet Backend for cBioPortal Web API

Doori Rose
M.S. Computer Science @ Georgia Tech

Mentors: Karthik Kalletla & Benjamin Gross

Why Spark & Parquet?

- Currently, cBioPortal uses MyBatis for the persistence layer and a relational database (MySQL) for data storage.
- The number and size of cancer datasets are expected to increase.
- **Spark**: a distributed computing engine for large-scale data processing.
  - Parallel processing faster than Hadoop
  - API - scala, Java, Python, R and SQL
- **Parquet** : a columnar storage format.

➜ Improve performance and user experience on large datasets

Spark Application Components

- Spark context configuration
- Utility for writing Parquet files
- Organization of Parquet files
- Performance
- Spark UI for monitoring the application

Spark Context Configuration

SparkContext - main component of Spark application that provides connection to the spark clusters/execution environment.

Name	Value	Name	Value
spark.app.name	cBioPortal	spark.master	local[*]
spark.default.parallelism	2	spark.scheduler.mode	FIFO
spark.driver.host	127.0.0.1	spark.sql.shuffle.partitions	2
spark.driver.memory	2g	spark.executor.memory	2g

Spark Configuration https://spark.apache.org/docs/latest/configuration.html

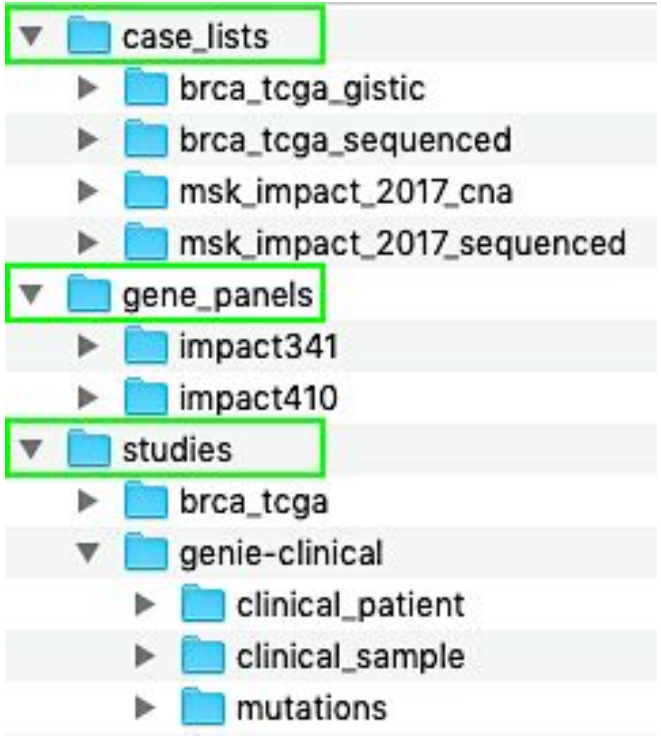
Parquet Writer Utility

ParquetWriter parameters:
--input-file : path to the TSV file
--output-file : path to write Parquet file
--type : type of file (case, panel, meta, cna, data by default)

$JAVA_HOME/bin/java -cp $HOME/cbioportal/scripts/target/scripts.jar org.cbioportal.persistence.spark.util.ParquetWriter* ***--input-file*** *$TSV_LOCATION/msk_impact_2017/data_clinical_sample.txt* ***--output-file*** *$PARQUET_DATA/studies/msk_impact_2017/clinical_sample* ***--type*** *data*

Organization of Parquet Files

TSV file	Parquet file
data_clinical_patient*.txt	clinical_patient
data_mutations*.txt	mutations
meta*.txt	meta
data_gene_panel_ impact341.txt	gene_panels/impact341
case_list/cases_cna.txt	case_lists/ msk_impact_2017_cna



Performance Results

URL / Request Body	Timing (ms)			
http://localhost:8080/api/filtered-samples/fetch	**SQL-10k**	**1k**	**10k**	**60k**
{"studyIds":["**msk_impact_2017**"]}	**2649.49**	1165.09	7703.30	33740.13
{"studyIds":["brca_tcga", "msk_impact_2017"]}	**2741.99**	15285.59	—	—
http://localhost:8080/api/sample-counts/fetch				
{“studyIds":["**msk_impact_2017**"]}	**1885.24**	7131.82	35201.71	n/a
{"studyIds":["brca_tcga", "msk_impact_2017"]}	**2213.38**	77325.68	—	—
http://localhost:8080/api/cna-genes/fetch				
{“studyIds":["**msk_impact_2017**"]}	**7595.85**	10924.89	16199.38	n/a
{"studyIds":["brca_tcga", "msk_impact_2017"]}	**21228.04**	31609.05	—	—

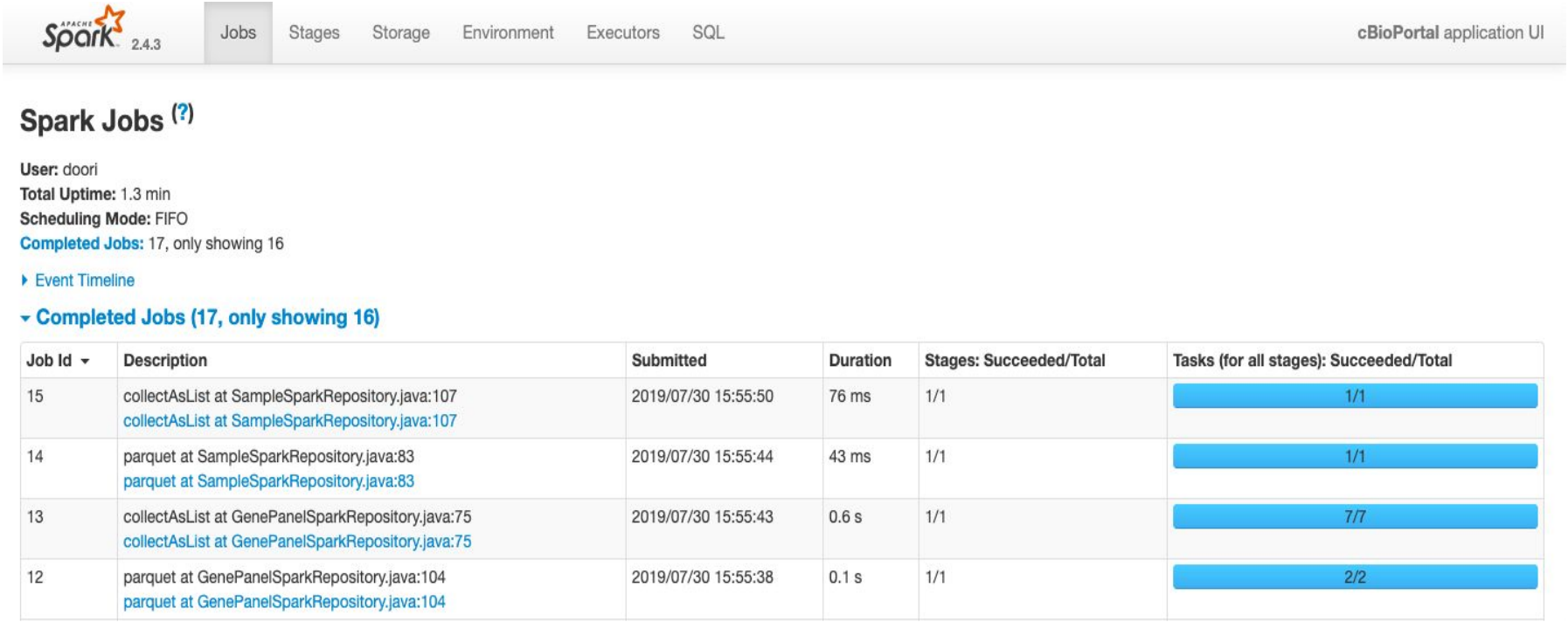
Performance Results

URL / Request Body	Timing (ms)			
http://localhost:8080/api/mutated-genes/fetch	SQL 10k	10k	SQL 60k	60k
{"studyIds":["msk_impact_2017"]}	936.28	4521.13	44676.29	40880.78

Big Data - 240k samples = 4x 60k clinical sample data & 2x mutation data

URL / Request Body	Timing (ms)	
http://localhost:8080/api/mutated-genes/fetch	SQL	Spark
{"studyIds":["genie-clinical"]}	365876.63	284925.50

Monitoring - Spark UI



Challenges and Future Improvements

- Integrating Spark/Parquet implementation with existing code written based on MySQL.
- Some data like cytoband were not readily available in data files.
- Spark & Parquet does not improve performance for all endpoints, however it can help with some APIs and large datasets.
- Performance can vary based on Spark configuration and the infrastructure.
- ParquetWriter can be extended to do more data clean up like the MySQL database import utilities.

Thank you!

Mentors: Benjamin & Karthik

Reference:

- Apache Spark https://spark.apache.org/
- Apache Spark Configuration https://spark.apache.org/docs/latest/configuration.html
- Apache Parquet https://parquet.apache.org/

Project description and code:

- GSoC project: https://summerofcode.withgoogle.com/projects/#5105508921376768
- Github spark-parquet-persistence branch: https://github.com/cBioPortal/cbioportal/tree/spark-parquet-persistence