Spark Parquet Implementation Performance

This document contains the performance analysis of the existing MyBatis implementation and Spark implementation for cBioPortal Study View APIs organized into three sections:

1. Spark Configuration
2. Study View Endpoints with Big Data
3. Study View Endpoints

1. Spark Configuration
A few Spark configuration settings were explored for my local environment to see their effect on performance. Spark configuration should be tweaked to fit the environment in which the Spark application is deployed.

http://localhost:8080/api/clinical-data-counts/fetch							
{"attributes":[{"attributeId":"CANCER_TYPE","clinicalDataType":"SAMPLE"}, {"attributeId":"CANCER_TYPE_DETAILED","clinicalDataType":"SAMPLE"}, {"attributeId":"SAMPLE_COUNT","clinicalDataType":"PATIENT"},{"attributeId":"SEX","clinicalDataType":"PATIENT"}, {"attributeId":"SAMPLE_TYPE","clinicalDataType":"SAMPLE"}],"studyViewFilter":{"studyIds":["msk_impact_2017"]}}							
spark.default.parallelism (4)		spark.executor.cores (4)		spark.sql.shuffle.partitions (200)		spark.driver.cores (1)	
1	16189.04	1	14396.25	1	14476.90	1	15421.01
2	**14460.93**	4	**13509.71**	**2**	**13463.79**	4	14852.12
4	15205.17	8	13966.97	100	23834.31	8	14795.29
8	15382.51	16	17124.62	200	33952.93	**16**	**14706.65**
driver, executor memory (g)							
1,1	16200.05						
1,2	**14772.22**						
2,1	15631.58						
2,2	14951.83						

Enabling dynamic allocation (spark.dynamicAllocation.enabled=true & spark.shuffle.service.enabled=true) does not affect the performance much, but could in a different environment or larger dataset.

Dynamic Allocation enabled: 14496.71 disabled: 14267.82
Dynamic Allocation enabled: 6475.92   disabled: 6440.73

2. Study View Endpoints with Big Data
In the next couple years, cBioPortal expects to see large datasets with up to 250k samples. A large mock dataset of 240k samples was created by doubling the mutation data and quadrupling the clinical sample

data of GENIE dataset. For *mutated-genes* api, Spark improves the timing when we use a dataset of 240k samples.

URL / Request Body	Timing (ms)	
	SQL	Spark
http://localhost:8080/api/mutated-genes/fetch		
{"studyIds":["genie-clinical"]}	197203.97	176397.10
http://localhost:8080/api/clinical-data-counts/fetch		
{"attributes": [{"attributeId":"ONCOTREE_CODE","clinicalDataType":"SAMPLE"}],"studyViewFilter": {"studyIds":["genie-clinical"]}}	6862.60	151688.00

3. Study View Endpoints
Performance of Study View endpoints were measured with 3 datasets: **brca_tcga** (1k), **msk_impact_2017** (10k), **genie** (60k). SQL-10k column serves as a baseline for the existing code, retrieving msk_impact_2017 dataset, and timings for multiple studies are also included.

URL / Request Body	Timing (ms)			
	SQL-10k	1k	10k	60k
http://localhost:8080/api/mutated-genes/fetch				
{"studyIds":["msk_impact_2017"]}	936.28	4521.13	1894.03	4071.24
{"studyIds":["brca_tcga", "msk_impact_2017"]}	11577.91	11539.89	—	—
http://localhost:8080/api/clinical-data-counts/fetch				
{"attributes": [{"attributeId":"CANCER_TYPE","clinicalDataType":"SAMPLE"}, {"attributeId":"CANCER_TYPE_DETAILED","clinicalDataType":"SAMPLE"}, {"attributeId":"SAMPLE_COUNT","clinicalDataType":"PATIENT"}, {"attributeId":"SEX","clinicalDataType":"PATIENT"}, {"attributeId":"SAMPLE_TYPE","clinicalDataType":"SAMPLE"}],"studyViewFilter":{"studyIds":["brca_tcga"]}}	804.09	2482.68	14470.17	78698.12
{"attributes": [{"attributeId":"CANCER_TYPE","clinicalDataType":"SAMPLE"}, {"attributeId":"CANCER_TYPE_DETAILED","clinicalDataType":"SAMPLE"}, {"attributeId":"SAMPLE_COUNT","clinicalDataType":"PATIENT"}, {"attributeId":"SEX","clinicalDataType":"PATIENT"}, {"attributeId":"SAMPLE_TYPE","clinicalDataType":"SAMPLE"}],"studyViewFilter":{"studyIds":["brca_tcga", "msk_impact_2017"]}}	1169.76	15317.67	—	—
{"attributes": [{"attributeId":"ONCOTREE_CODE","clinicalDataType":"SAMPLE"}],"studyViewFilter":{"studyIds":["genie-clinical-60k"]}}	103.99	807.09	5413.32	27947.21

URL / Request Body	Timing (ms)			
	SQL-10k	1k	10k	60k
http://localhost:8080/api/clinical-data-bin-counts/fetch?dataBinMethod=STATIC				
{"attributes": [{"attributeId":"MUTATION_COUNT","clinicalDataType":"SAMPLE","disableLogScale":false}, {"attributeId":"FRACTION_GENOME_ALTERED","clinicalDataType":"SAMPLE","disableLogScale":false}, {"attributeId":"DNA_INPUT","clinicalDataType":"SAMPLE","disableLogScale":false}, {"attributeId":"OS_MONTHS","clinicalDataType":"PATIENT","disableLogScale":false}, {"attributeId":"SAMPLE_COVERAGE","clinicalDataType":"SAMPLE","disableLogScale":false}],"studyViewFilter":{"studyIds":["msk_impact_2017"]}} {"attributes": [{"attributeId":"MUTATION_COUNT","clinicalDataType":"SAMPLE","disableLogScale":false}, {"attributeId":"FRACTION_GENOME_ALTERED","clinicalDataType":"SAMPLE","disableLogScale":false}, {"attributeId":"AGE","clinicalDataType":"PATIENT","disableLogScale":false}, {"attributeId":"DAYS_TO_COLLECTION","clinicalDataType":"SAMPLE","disableLogScale":false}],"studyViewFilter":{"studyIds":["**brca_tcga**"]}}	1736.35	8185.59	57280.60	n/a
{"attributes": [{"attributeId":"MUTATION_COUNT","clinicalDataType":"SAMPLE","disableLogScale":false}, {"attributeId":"FRACTION_GENOME_ALTERED","clinicalDataType":"SAMPLE","disableLogScale":false}], "studyViewFilter":{"studyIds":["msk_impact_2017", "brca_tcga"]}}	1214.92	18410.37	—	—
http://localhost:8080/api/clinical-data-density-plot/fetch?xAxisAttributeId=FRACTION_GENOME_ALTERED&xAxisBinCount=50&xAxisStart=0&xAxisEnd=1&yAxisAttributeId=MUTATION_COUNT&yAxisBinCount=52&yAxisStart=0&clinicalDataType=SAMPLE				
{"studyIds":["**msk_impact_2017**"]}	285.4	3145.00	21699.01	n/a
{"studyIds":["brca_tcga", "msk_impact_2017"]}	776.17	34560.96	—	—
http://localhost:8080/api/filtered-samples/fetch				
{"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				

URL / Request Body	Timing (ms)			
	SQL-10k	1k	10k	60k
{"studyIds":["**msk_impact_2017**"]}	7595.85	10924.89	16199.38	n/a
{"studyIds":["brca_tcga", "msk_impact_2017"]}	21228.04	31609.05	—	—
http://localhost:8080/api/copy-number-enrichments/fetch? copyNumberEventType=HOMDEL&enrichmentType=PATIENT				
[{"name":"(A) Metastasis","molecularProfileCaseIdentifiers": [{"caseId":"P-0000197","molecularProfileId":"msk_i mpact_2017_cna"},...]}, {"name":"(B) Primary","molecularProfileCaseIdentifiers": [{"caseId":"P-0000377","molecularProfileId":"msk_i mpact_2017_cna"},...]}]		—	4077.54	—
http://localhost:8080/api/copy-number-enrichments/fetch? copyNumberEventType=AMP&enrichmentType=PATIENT				
[{"name":"(A) Metastasis","molecularProfileCaseIdentifiers": [{"caseId":"P-0000197","molecularProfileId":"msk_i mpact_2017_cna"},...]}, {"name":"(B) Primary","molecularProfileCaseIdentifiers": [{"caseId":"P-0000377","molecularProfileId":"msk_i mpact_2017_cna"}]}]		—	3190.12	—