Hadoop in practice 2 – Notes

https://github.com/aspdave Dave Jan2016

1. Background and fundamentals, Hadoop in a heartbeat, what is Hadoop? Core Hadoop components, The Hadoop ecosystem, Hardware requirements, Hadoop distributions, Who's using Hadoop? , Hadoop limitations, getting your hands dirty with MapReduce, summary

2. Introduction to YARN, YARN overview, Why YARN? YARN concepts and components, YARN configuration, Interacting with YARN, YARN challenges, YARN and MapReduce, Dissecting a YARN MapReduce application, Configuration, Backward compatibility, Running a job, Monitoring running jobs and viewing archived jobs, Uber jobs, YARN applications, NoSQL, Interactive SQL, Graph processing, eel-time data processing, Bulk synchronous parallel, MPI, In-memory, DAG execution, Data logistics.

3. Data serialization—working with text and beyond, Understanding inputs and outputs in MapReduce, Data input, Data output, Processing common serialization formats, XML, JSON, Big data serialization formats, Comparing Sequence File, Protocol Buffers, Thrift, and Avro, Sequence File, Protocol Buffers, Thrift, Avro,Columnar storage, Understanding object models and storage formats, Parquet and the Hadoop ecosystem, Parquet block and page sizes, Parquet limitations, Custom file formats, Input and output formats, The importance of output committing, summary.

4. Organizing and optimizing data in HDFS, Data organization, Directory and file layout, Data tiers, Partitioning, Compacting, Atomic data movement, efficient storage with compression, summary

5. Moving data into and out of Hadoop , Key elements of data movement, Idempotence, Aggregation, Data format transformation, Compression, Availability and recoverability, Reliable data transfer and data validation, Resource consumption and performance, Monitoring ,Speculative execution ,Moving data into Hadoop, Roll your own ingest, Continuous movement of log and binary files into HDFS , Databases, 5.2.4. Base, Importing data from Kafka, Moving data out of Hadoop, Roll your own egress , Automated file egress, Databases, NoSQL, Big data patterns ,summary

Technique 1 Determining the configuration of your cluster

Technique 2 Running a command on your YARN cluster

Technique 3 Accessing container logs

Technique 4 Aggregating container log files

Technique 5 Writing code that works on Hadoop versions 1 and 2

Technique 6 Using the command line to run a job

Technique 7 Running small MapReduce jobs

Technique 8 MapReduce and XML

Technique 9 MapReduce and JSON

Technique 10 Working with SequenceFiles

Technique 40 An automated mechanism to copy files into HDFS

Technique 41 Scheduling regular ingress activities with Oozie

Technique 42 Using Sqoop to import data from MySQL

Technique 43 HBase ingress into HDFS

Technique 44 MapReduce with HBase as a data source

Technique 45 Using Camus to copy Avro data from Kafka into HDFS

Technique 46 Using the CLI to extract files

Technique 47 Using REST to extract files

Technique 48 Reading from HDFS when behind a firewall

Technique 49 Mounting Hadoop with NFS

Technique 50 Using DistCp to copy data out of Hadoop

Technique 51 Using Java to extract files

Technique 52 An automated mechanism to export files from HDFS

Technique 53 Using Sqoop to export data to MySQL

6. Applying MapReduce patterns to big data, Joining, Join data, Map-side joins, Reduce-side joins, Data skew in reduce-side joins, Sorting, Secondary sort, Total order sorting, Sampling,Summary.

7. Utilizing data structures and algorithms at scale, Modeling data and solving problems with graphs, Modeling graphs, Shortest-path algorithm, Friends-of-friends algorithm, Using Giraph to calculate PageRank over a web graph, Bloom filters, HyperLogLog, A brief introduction to HyperLogLog,summary

8. Tuning, debugging, and testing, Measure, measure, measure, Tuning MapReduce, Common inefficiencies in MapReduce jobs, Map optimizations, Shuffle optimizations, Reducer optimizations, General tuning tips, Debugging, Accessing container log output, Accessing container start scripts, Debugging OutOfMemory errors, MapReduce coding guidelines for effective debugging, Testing MapReduce jobs, Essential ingredients for effective unit testing, MRUnit,LocalJobRunner, MiniMRYarnCluster,Integration and QA testing, Beyond MapReduce ,summary

9. SQL on Hadoop, Hive, Hive basics, Reading and writing data ,User-defined functions in Hive, Hive performance, Impala,Impala vs. Hive,Impala basics, User-defined functions in Impala, Spark SQL,Spark 101,Spark on Hadoop,SQL with Spark,summary

10. Writing a YARN application,Fundamentals of building a YARN application,Actors,The mechanics of a YARN application, Building a YARN application to collect cluster statistics, Additional YARN application capabilities,RPC between components,Service discovery,Checkpointing application progress,Avoiding split-brain,Long-running applications, Security,YARN programming abstractions,Twill,Spring,REEF , Picking a YARN API abstraction, summary .

Technique 80 Using stack dumps to discover unoptimized user code

Technique 81 Profiling your map and reduce tasks

Technique 82 Examining task logs

Technique 83 Figuring out the container startup command

Technique 84 Force container JVMs to generate a heap dump

Technique 85 Augmenting MapReduce code for better debugging

Technique 86 Using MRUnit to unit-test MapReduce

Technique 87 Heavyweight job testing with the LocalJobRunner

Technique 88 Using MiniMRYarnCluster to test your jobs

Technique 89 Working with text files

Technique 90 Exporting data to local disk

Technique 91 Writing UDFs

Technique 92 Partitioning

Technique 93 Tuning Hive joins

Technique 94 Working with text

Technique 95 Working with Parquet

Technique 96 Refreshing metadata

Technique 97 Executing Hive UDFs in Impala

Technique 98 Calculating stock averages with Spark SQL

Technique 99 Language-integrated queries

Technique 100 Hive and Spark SQL

Technique 101 A bare-bones YARN client

Technique 102 A bare-bones ApplicationMaster

Technique 103 Running the application and accessing logs

Technique 104 Debugging using an unmanaged application master

==========