Hadoop Documentation (L2)

1. MapReduce Types and Formats
   1. MapReduce Types
   2. Input Formats
   3. Output Formats
2. MapReduce Features

2.1 Counters

2.2 Sorting

2.3 Joins

2.4 Side Data Distribution

2.5 MapReduce Library Classes

PIG L2

1. Installing and Running Pig
2. An Example
3. Comparison with Databases
4. Pig Latin
5. User\_ Defined Functions
6. Data Processing Operators
7. Pig in practice

SQOOP L2

1. Importing Data
   1. Transferring an Entire Table
   2. Specifying a target Directory
   3. Importing only a subset of data
   4. Protecting your password
   5. Using a file format other than CSV
   6. Compressing imported Data
   7. Speeding up Transfers
   8. Overriding Type mapping
   9. Controlling parallelism
   10. Encoding NULL Values
   11. Importing all your tables
   12. Incremental import
   13. Free-form Query import
2. Export

2.1 Transferring data from Hadoop

2.2 Inserting Data in Batches

2.3 Exporting with All –or-nothing semantics

2.4 updating an Existing Data set

2.5 Updating or inserting at the same time

2.6 Using stored procedures

2.7 Exporting into a subset of Columns

2.8 Encoding the NULL Value Differently

2.9 Exporting Corrupted Data

1. Hadoop Ecosystem Integration

3.1 Scheduling sqoop jobs with Oozie

3.2 specifying Commands in Oozie

3.3 Using property parameters in Oozie

3.4 Installing JDBC Drivers in Oozie

3.5 Importing data directly into Hive

3.6 Using Partitioned Hive tables

3.7 Replacing special Delimiters During Hive Import

3.8 Using the Correct Null String in Hive

3.9 Importing Data into Hbase

3.10 Importing All Rows into Hbase

3.11 Improving performance when importing into HBase

1. Connectors

4.1 Overriding Imported Boolean Values in PostgreSQL Direct Import

4.2 Importing a Table stored in Custom Schema in PostgreSQL

4.3 Exporting into PostgreSQL using pg\_bulkload

4.4 Connecting to MYSQL

4.5 Using Direct MYSQL Import into Hive

4.6 Using the upsert feature when exporting into MYSQL

4.7 Importing from Orale

4.8 using synonyms in Oracle

4.9 Faster Transfers with Oracle

4.10 importing into Avro with oraOop

4.11 Choosing the proper Connector for Oracle

4.12 Exporting into Teradata

4.13 Using the Cloudera Teradata Connector

4.14 Using Long Column Names in Teradata

HIVE L2

1. HiveQL Indexes
   1. Creating an index
   2. Rebuilding the index
   3. Showing an index
   4. Dropping an index
   5. Implementing a Custom index Handler
2. Schema Design

2.1 Table –by-Day

2.2 Over Partitioning

2.3 Unique keys and Normalization

2.4 Making Multiple passes over the same Data

2.5 The Case for Partitioning Every Table

2.6 Bucketing table data storage

2.7 Adding Columns to a Table

2.8 Using Columnar Tables

1. Partitions and Buckets

3.1 Storage Formats

1. Tuning

4.1 Using Explain

4.2 EXPLAIN EXTENDED

4.3 Limit Tuning

4.4 Optimized joins

4.5 Local Mode

4.6 Parallel Execution

4.7 Strict Mode

4.8 Tuning the number of Mappers and Reducers

4.10 Indexes

4.11 Dynamic Partition Tuning

4.12 Speculative Execution

4.13 single MapReduce MultiGROUP BY

4.14 Virtual Columns

1. File Formats and Compression

5.1 Determining installed codecs

5.2 Choosing a Compression Codec

5.3 Enabling intermediate Compression

5.4 Final output compression

5.5 sequence files, ORC File

5.6 Compression in Action

5.7 Archive parttion

5.8 Compression: Wrapping up

1. Log Properties

6.1 Changing Log4j properties

6.2 connecting a java Debugger to hive

6.6 unit Testing in hive with hive\_ test

6.7 The new plugin Developer Kit

1. Hive Functions

7.1 Discovering and Describing Functions

7.2 Calling Functions

7.3 Standard Functions

7.4 Aggregate Functions

7.5 Table Generating Functions

7.6 A UDF for finding a Zodiac sign from a Day

7.7 UDF Versus GenericUDF

7.8 Permanent Functions

7.9 User –Defined Aggregate functions

7.10 User- Defined Table generating Functions

7.11 Accessing the Distributed cache from a UDF

7.12 Annotations for use with functions

7.13 Macros

1. Streaming(Overview)

8.1 Identity Transformation

8.2 Changing Types

8.3 Projecting Transformation

8.4 Manipulative Transformations

8.5 Using the Distributed Cache

8.6 Producing Multiple rows from a single row

8.7 Calculating Aggregates with Streaming

8.8 CLUSTER BY, DISTRIBUTE BY , SORT BY

8.9 GenericMR Tools for Streaming to java

8.10 Calculating Cogroups

1. Customizing Hive File and Record Formats

9.1 File Versus Record formats

9.2 Demystifying CREATE TABLE Statements

9.3 File Formats

9.4 Record Formats: SerDes

9.5 CSV and TSV SerDes

9.6 Objectinspector

9.7 XML UDF

9.8 Xpath-Related Functions

9.9 JSON SerDe

9.10 Avro Hive SerDe

9.11 Binary Output

1. Hive Thrift Service(Overview)

10.1 Starting the Thrift server

10.3 connecting to hive server

10.4 getting cluster status

10.5 Result set Schema

10.6 Fetching Results

10.7 Retrieving Query plan

1. Hcatalog (Overview)

11.1 Introduction

11.2 Mapreduce

11.3 Reading Data

11.4 Writing Data

11.5 Command Line

11.6 Security Model

11.7 Architecture