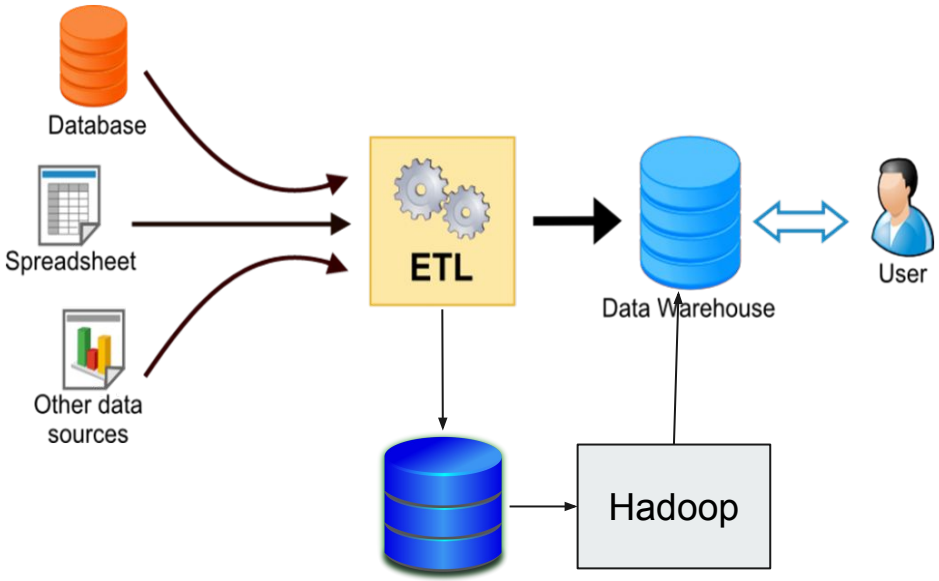
Hadoop ETL-ELT

- Abhay Dandekar

Agenda

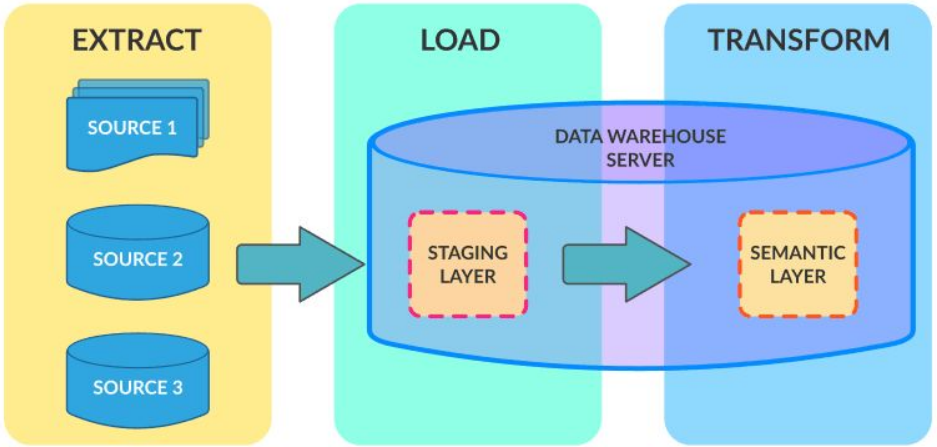
1. What is ETL?
2. What is ELT ?
3. Tools for ETL
4. Pig
5. Sqoop

ETL - Extract Transform Load



1. Homegeneous / Heterogeneous data sources
2. Data is transformed in proper format.
    a. Joins
    b. Lookups
    c. Aggregate (rollups etc)
    d. Transposing
    e. Pivoting
3. Data is loaded into final target database.
    a. HDFS
    b. SQL

ELT - Extract Load Transform



1. Homegeneous / Heterogeneous data sources
2. Data is loaded into interim warehouses / databases layers.
    a. Hive / HBase
3. Business logic is applied over data, transformed in proper format and kept into a consumer layer.

Create table DQ_GOOD.table1 as

Select * from stage.table1

Where primary_key1 is not null;

Pig

1. Created at Yahoo!
2. Data flow language
3. Language to express data-flows, Pig Latin
4. Pig has two execution environments
    a. Local (Single JVM, accesses local path , pig -x local)
    b. Distributed execution on Hadoop cluster (Clustered, default mode)
5. Pig turns your program into a series of MR jobs

Pig Latin

1. Statements must be terminated by semi-colon ;
2. LOAD - Load a file
3. AS - provide the schema for file
4. DUMP - dump data on screen
5. JOIN - Like for SQL queries
6. $0 / $1 - Can be used to load columns 0, 1 etc

Commands

Table 16-1. Pig Latin relational operators

Category	Operator	Description
Loading and storing	`LOAD`	Loads data from the filesystem or other storage into a relation
	`STORE`	Saves a relation to the filesystem or other storage
	`DUMP (\d)`	Prints a relation to the console
Filtering	`FILTER`	Removes unwanted rows from a relation
	`DISTINCT`	Removes duplicate rows from a relation
	`FOREACH… GENERATE`	Adds or removes fields to or from a relation
	`MAPREDUCE`	Runs a MapReduce job using a relation as input
	`STREAM`	Transforms a relation using an external program
	`SAMPLE`	Selects a random sample of a relation
	`ASSERT`	Ensures a condition is true for all rows in a relation; otherwise, fails
Grouping and joining	`JOIN`	Joins two or more relations
	`COGROUP`	Groups the data in two or more relations
	`GROUP`	Groups the data in a single relation
	`CROSS`	Creates the cross product of two or more relations
	`CUBE`	Creates aggregations for all combinations of specified columns in a relation
Sorting	`ORDER`	Sorts a relation by one or more fields
	`RANK`	Assign a rank to each tuple in a relation, optionally sorting by fields first
	`LIMIT`	Limits the size of a relation to a maximum number of tuples
Combining and splitting	`UNION`	Combines two or more relations into one
	`SPLIT`	Splits a relation into two or more relations

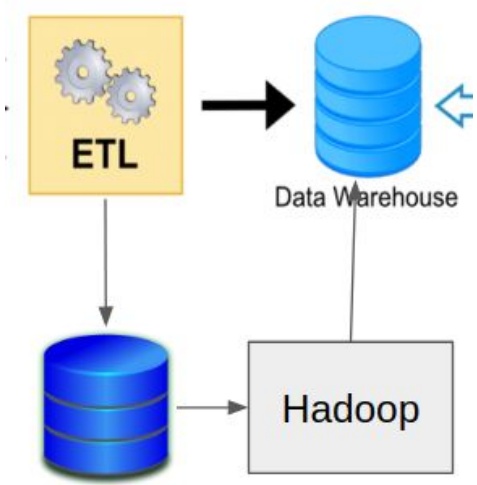
Further reading

1. https://pig.apache.org/docs/latest/index.html
2. Hadoop - A definitive guide, 4th Edition, Chapter 16

SQOOP - LOAD UNLOAD from DBs

Sqoop

1. Allows to extract data from RDBMS into Hadoop for further processing
2. Once the data is processed, it can be pushed back into RDBMS for further consumption by users
3. Sqoop already ships will connectors with popular DBs like MySQL, Netezza, Postgres, Oracle etc

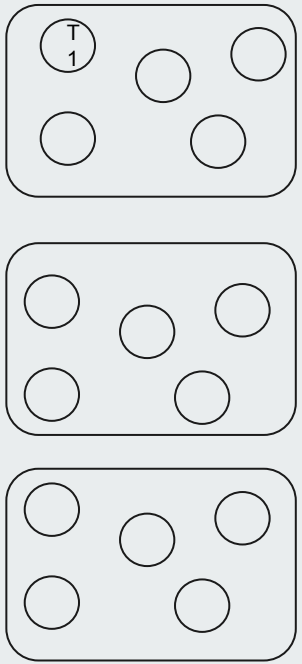


Execution

1. Import
2. Codegen
3. Import All Tables
4. Import with where clause
5. Export

Sqoop Import / Export internals

1. Sqoop connects to DB and verifies the Table schema
2. Gets a list of all columns and their SQL types
3. SQL types are mapped to Java Data types
4. Sqoops codegen will generate the code
5. Code gets executed as Mapper / Reducer




```
Select * from schema1.t1 ;
Select * from schema1.t2;

Select * from schema2.t1;
Select * from schema2.t100;
```

Further reading

1. http://sqoop.apache.org/docs/1.4.5/SqoopUserGuide.html
2. Hadoop: Definitive Guide, 4th Edition Chapter 15