Big Data
Hive

Dr. Wenceslao PALMA
wenceslao.palma@pucv.cl

Hive: Intro

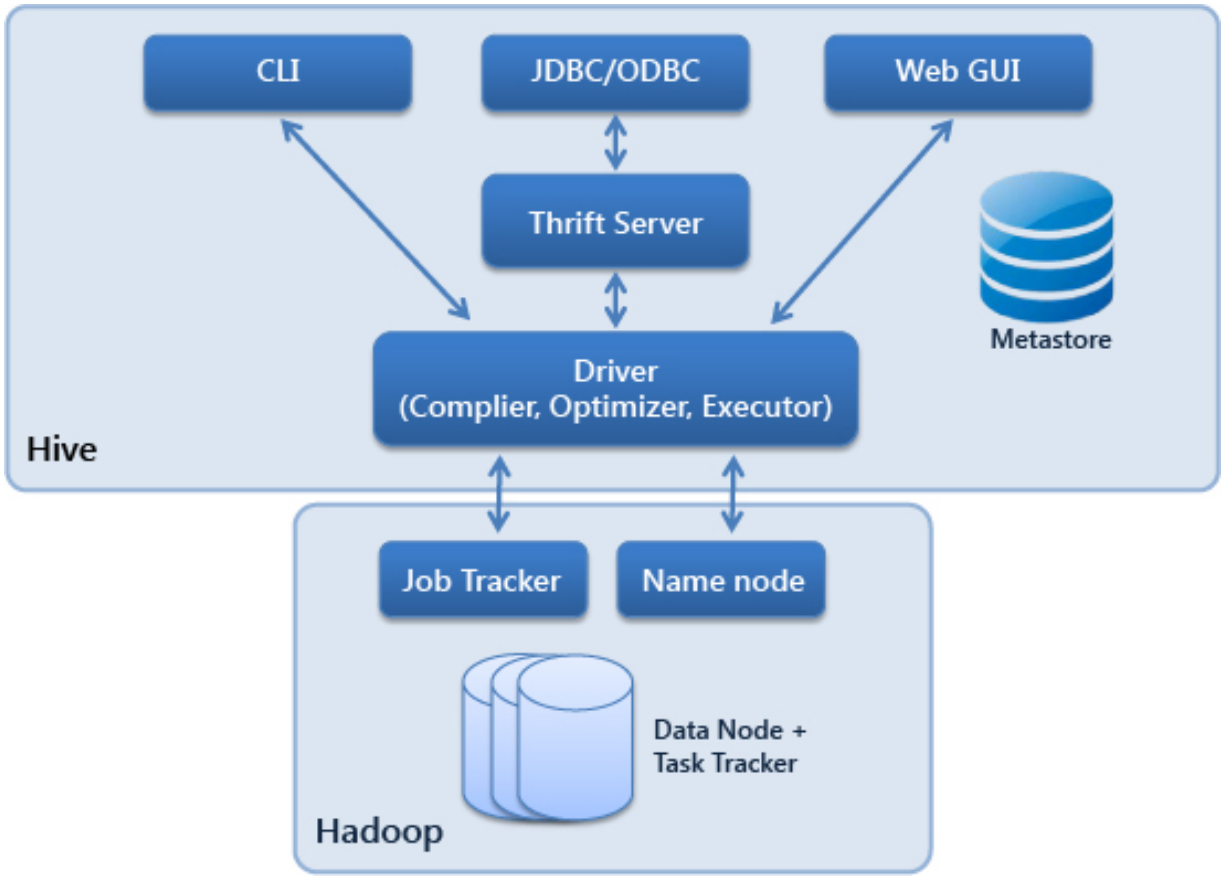
- Is a data warehousing infrastructure based on Hapache Hadoop.
- Is designed to enable.
  - Easy data summarization.
  - Ad-hoc querying and analysis of large volumes of data using a SQL-like language called HiveQL.
  - Integrate custom analysis through UDFs.
- Uses map-reduce for execution.
- Open data formats (parquet, avro, apache ORC).
- HDFS for storage.

Hive: Intro

- Is a data warehousing infrastructure based on Hapache Hadoop.
- Is designed to enable.
  - Easy data summarization.
  - Ad-hoc querying and analysis of large volumes of data using a SQL-like language called HiveQL.
  - Integrate custom analysis through UDFs.
- Uses map-reduce for execution.
- Open data formats (parquet, avro, apache ORC).
- HDFS for storage.

Hive is not designed for online transaction processing.

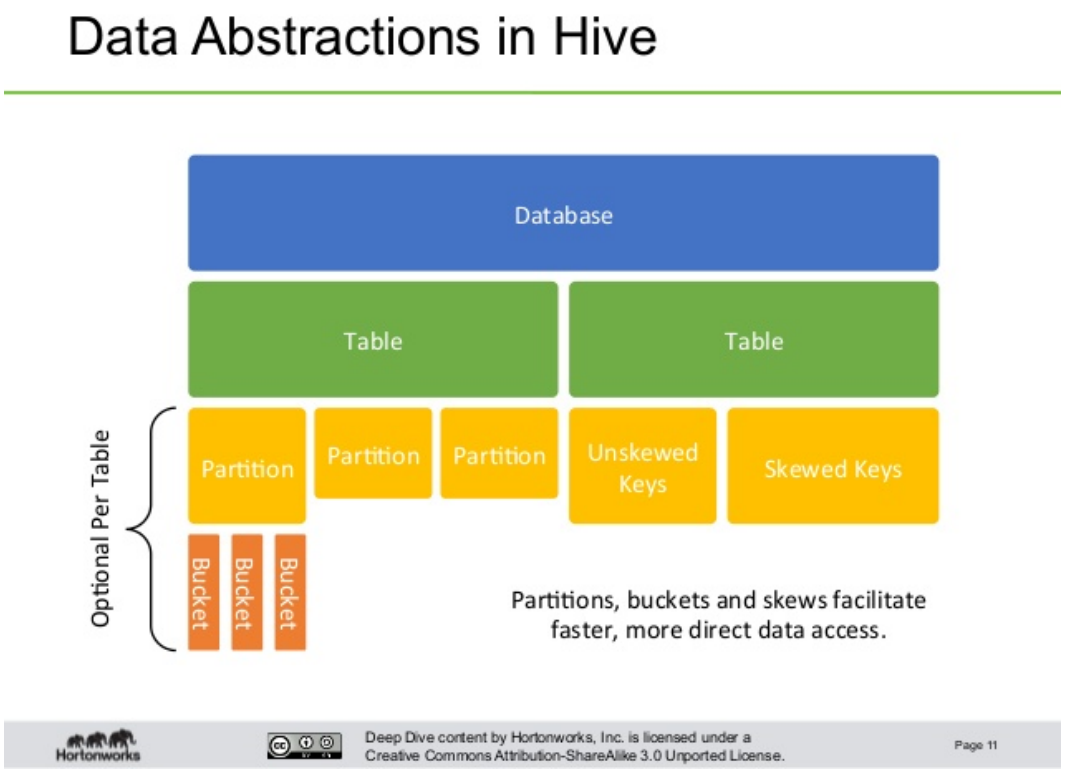
Hive: Components



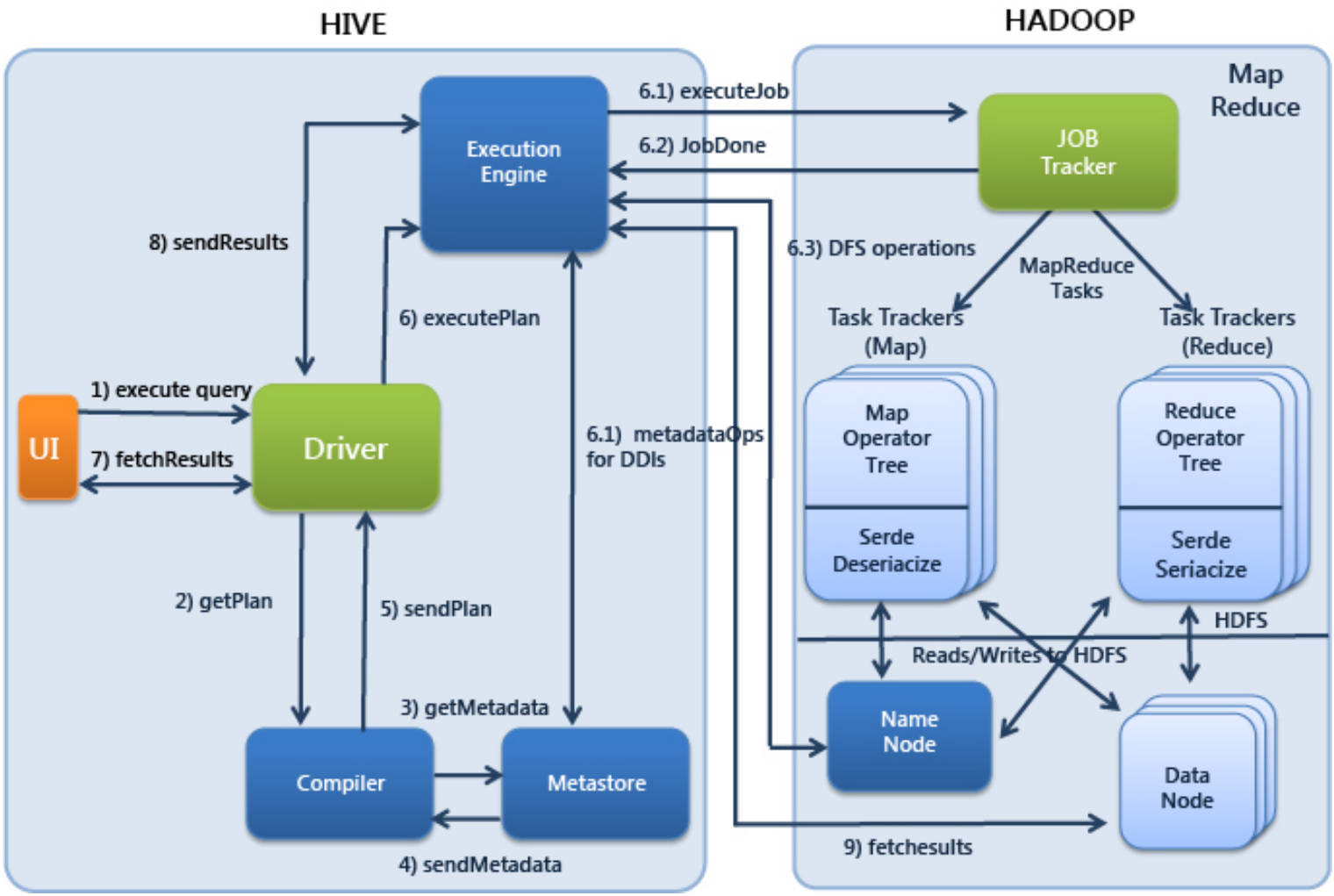
Hive: Data Abstractions

Hive data is organized into:

- Databases
- Tables
- Partitions
- Buckets (Clusters)



Hive: mapreduce



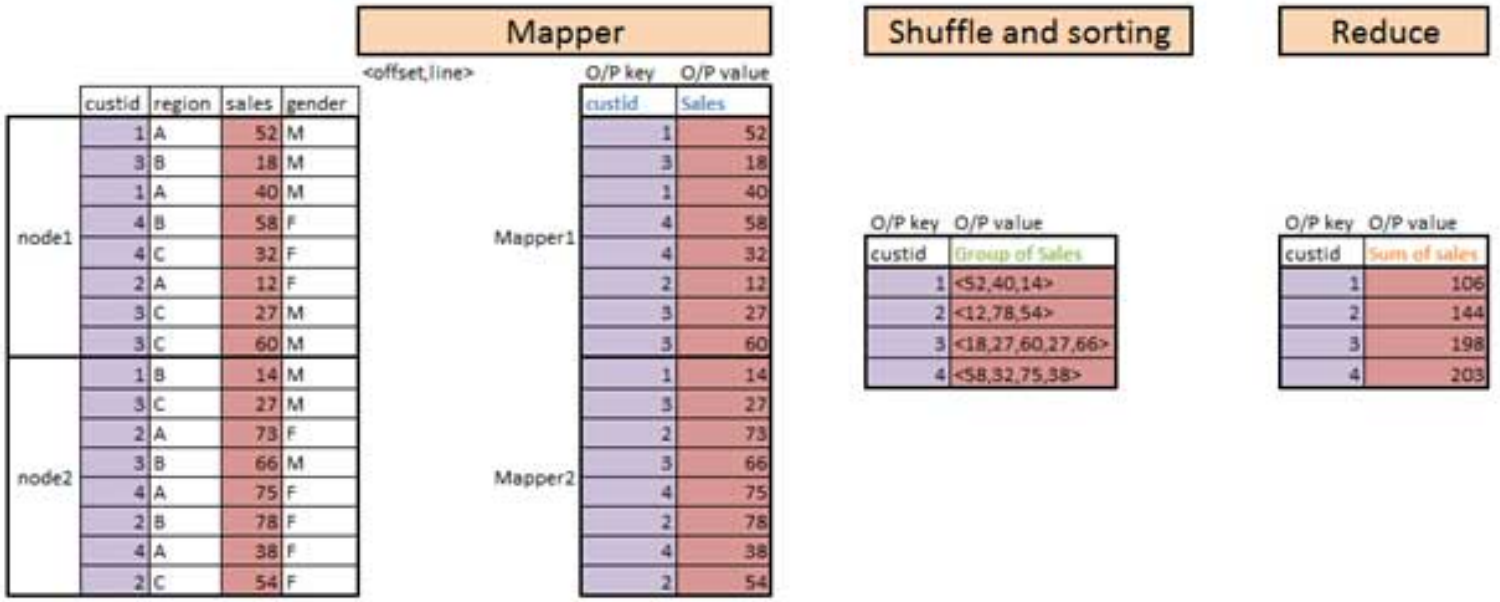
Hive: How it is mapped to mapreduce [1]

	custid	region	sales	gender
node1	1	A	52	M
	3	B	18	M
	1	A	40	M
	4	B	58	F
	4	C	32	F
	2	A	12	F
	3	C	27	M
	3	C	60	M
node2	1	B	14	M
	3	C	27	M
	2	A	73	F
	3	B	66	M
	4	A	75	F
	2	B	78	F
	4	A	38	F
	2	C	54	F

- data is organized in 4 fields.
- data is stored and distributed using HDFS.
- Query: find the customer total sales.

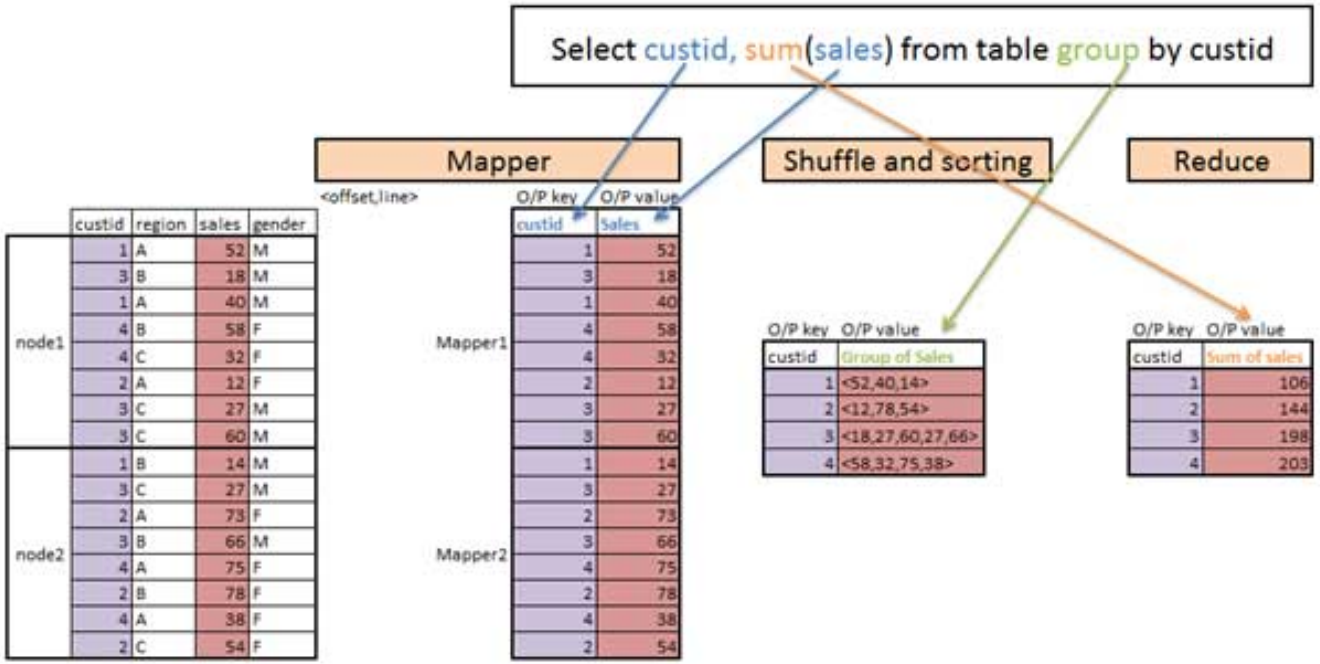
[1] images from http://www.edupristine.com/blog/facebook-hive-explained

Hive: (1) Find the customer total sales using mapreduce



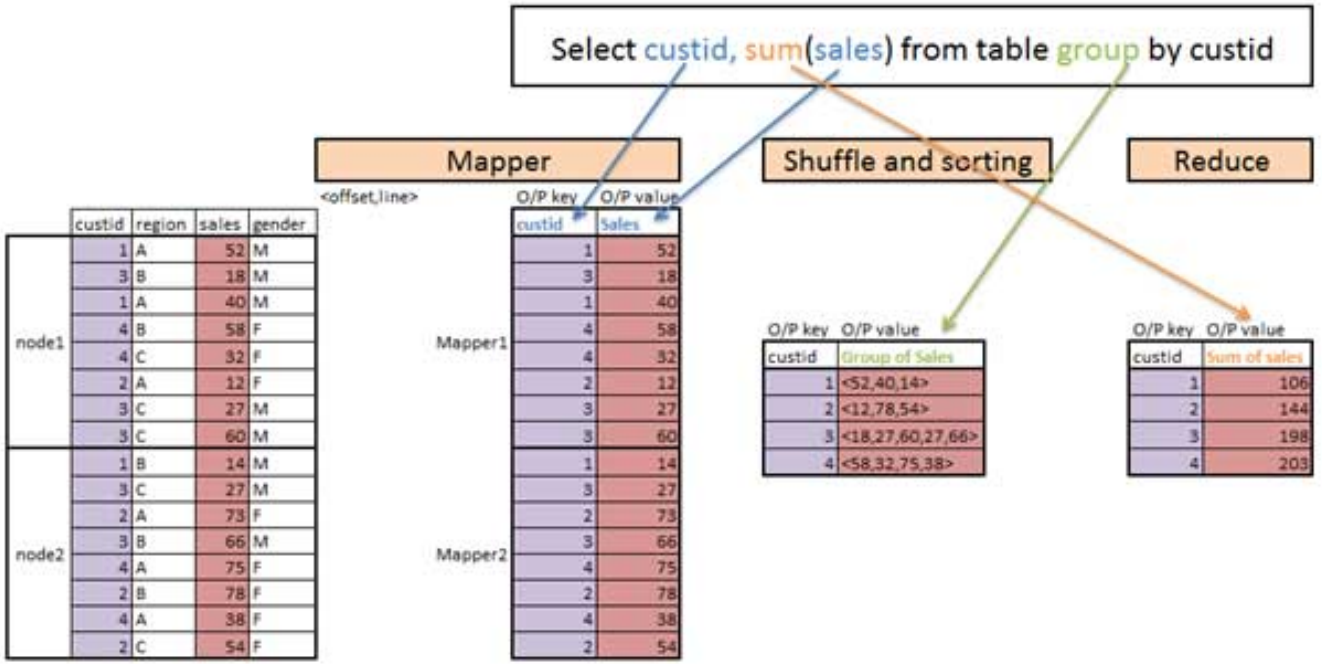
- How to tackle it using mapreduce?
- In the map phase: $< key, value > = < custid, sales >$
- In the reduce phase, the total sum of sales is calculated for each *custid*.

Hive: (1) Find the customer total sales using Hive



- $< key, value >$ is taken from the Select clause.
- the group by clause happens in Shuffle.
- the aggregation happens in reduce.

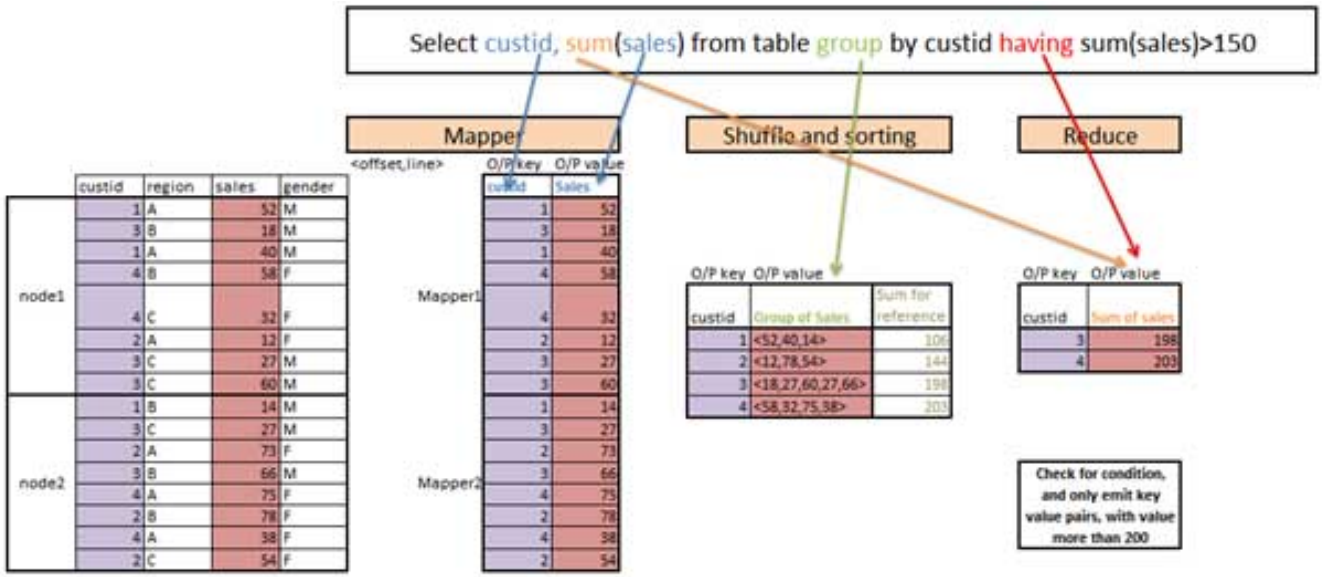
Hive: (1) Find the customer total sales using Hive



- $< key, value >$ is taken from the Select clause.
- the group by clause happens in Shuffle.
- the aggregation happens in reduce.

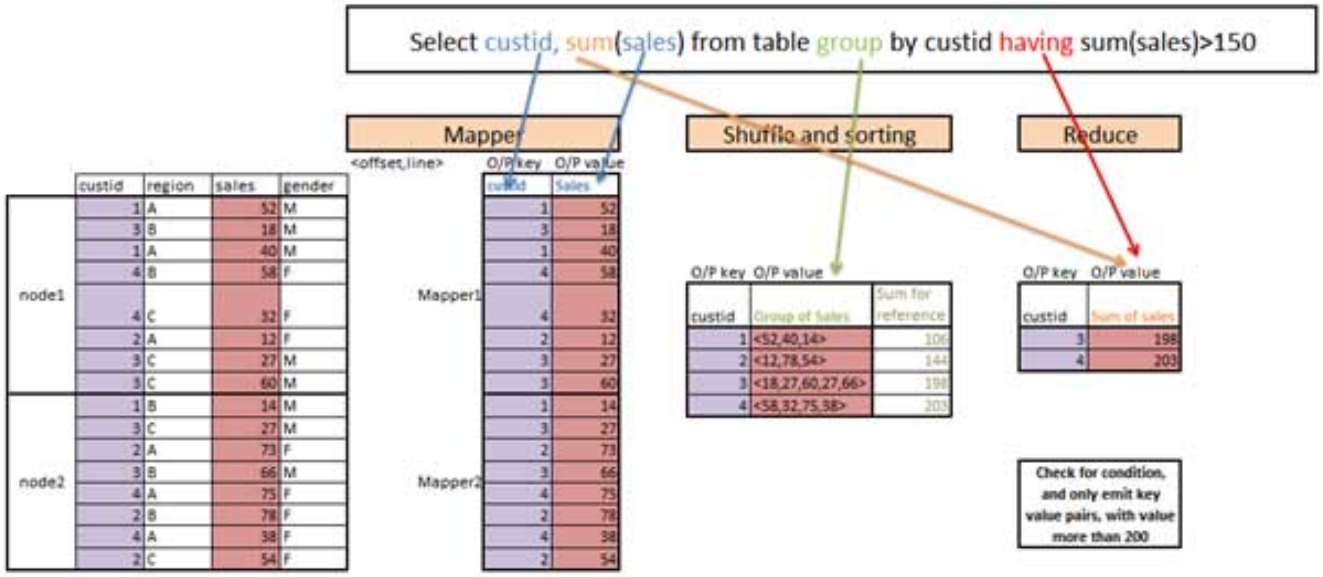
Data analysts/scientists focus on understand data and bring quick insights

Hive: (2) Identify customers whom a credit card can be offered using mapreduce



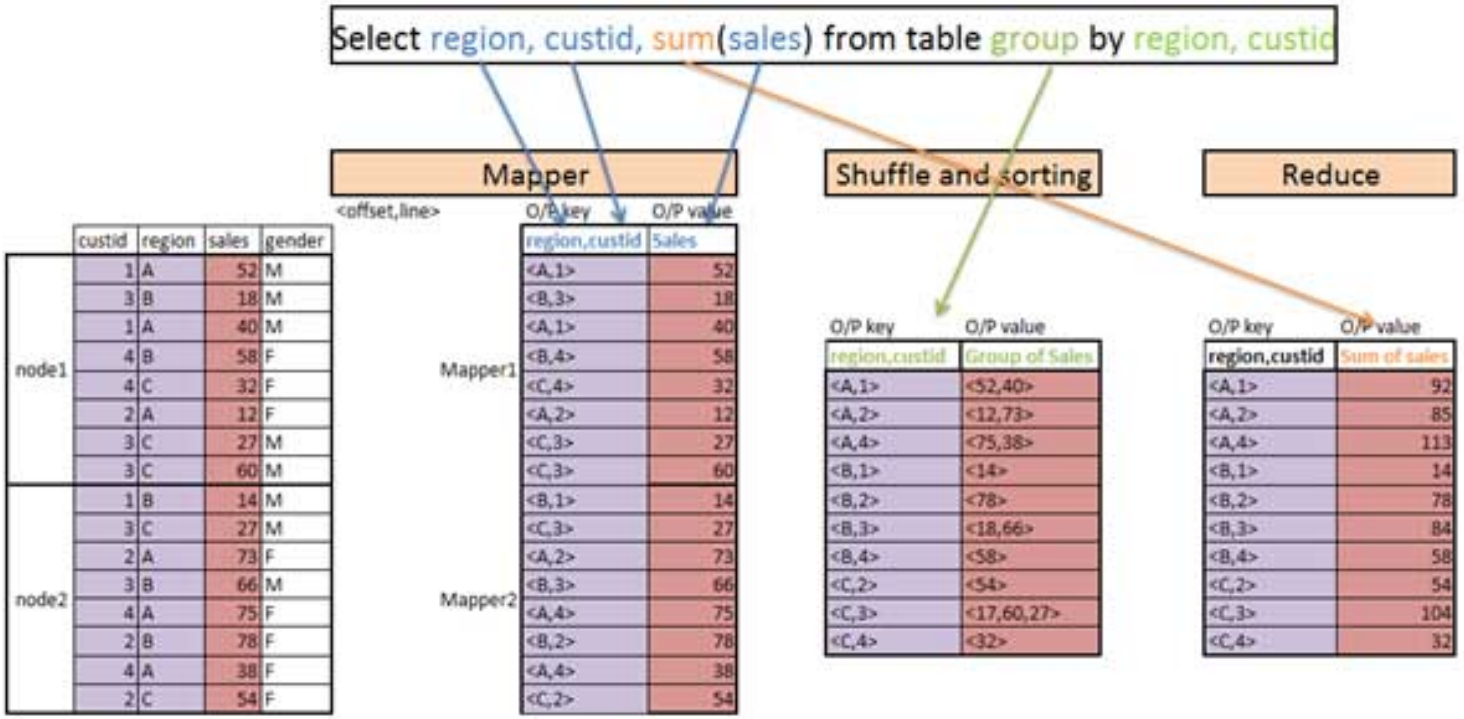
- In the map phase: $< key, value > = < custid, sales >$
- In the shuffle phase data is grouped using $custid$
- In the reduce phase, all the customers with $sum(sales) > 150$ are filtered.

Hive: (2) Identify customers whom a credit card can be offered using HIVE



- $< key, value >$ is taken from the Select clause.
- the group by happens in shuffle.
- aggregation happens in reduce.

Hive: (3) Calculate customers spendings across differents regions



Hive: Mapping functionalities from Hive to MapReduce

Query	Map Reduce Stage
Selection	Mapper
Where	Mapper
group	Shuffle
Row level functions	Mapper
Case statements, if statements	Mapper
Field level functions, like **UPPER**	Mapper
Order	Sort
Group level aggregations like avg, sum, max	Reduce
Having	Reduce