Introduction to Hive

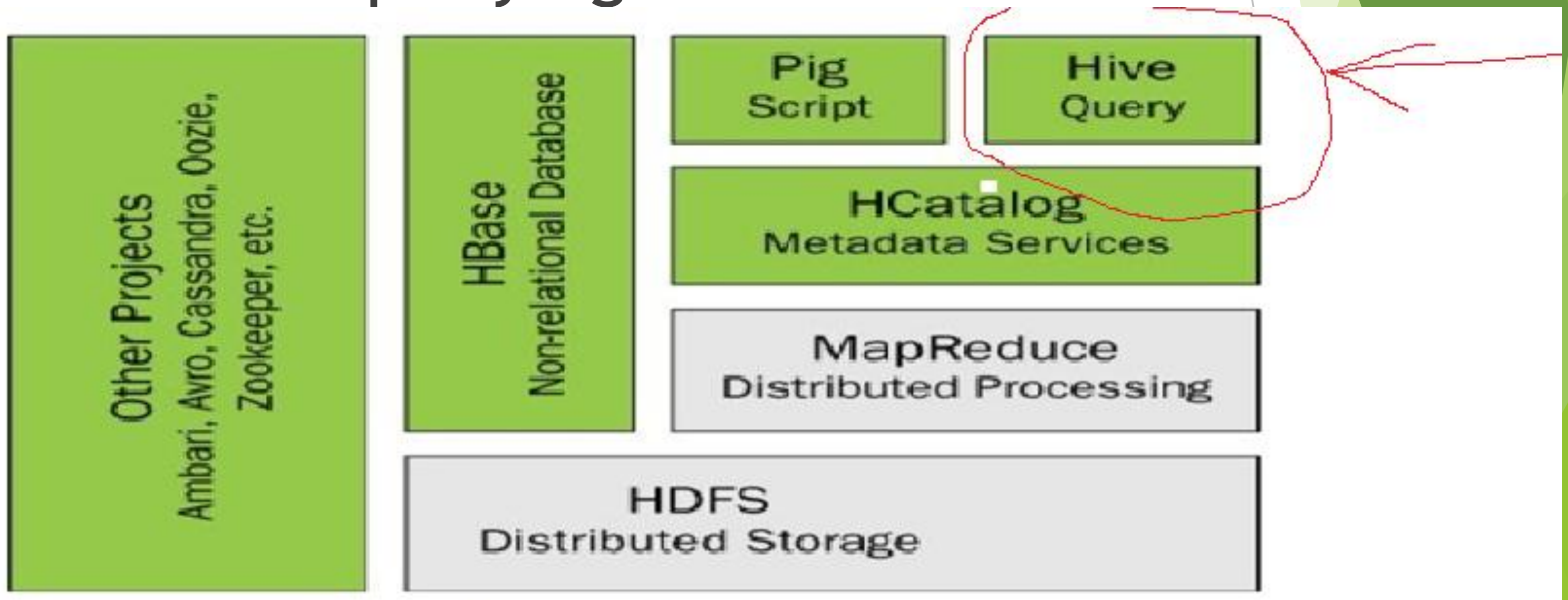
Tuhin Mahmud
@Big Data Revealed Meetup group
10/11/2014

About me

- Name: Tuhin Mahmud
- Software developer @IBM
- Cloudera Certified Hadoop Developer
- Email:tuhinm@hotmail.com
- www.linkedin.com/in/tuhinmahmud

What is Hive?

- Hive is an essential Hadoop ecosystem tool that provides an SQL dialect for querying data stored in Hadoop.[1]



The Hadoop 1.0 ecosystem.

Arun C Murthy , Vinod Kumar Vavilapalli[3]

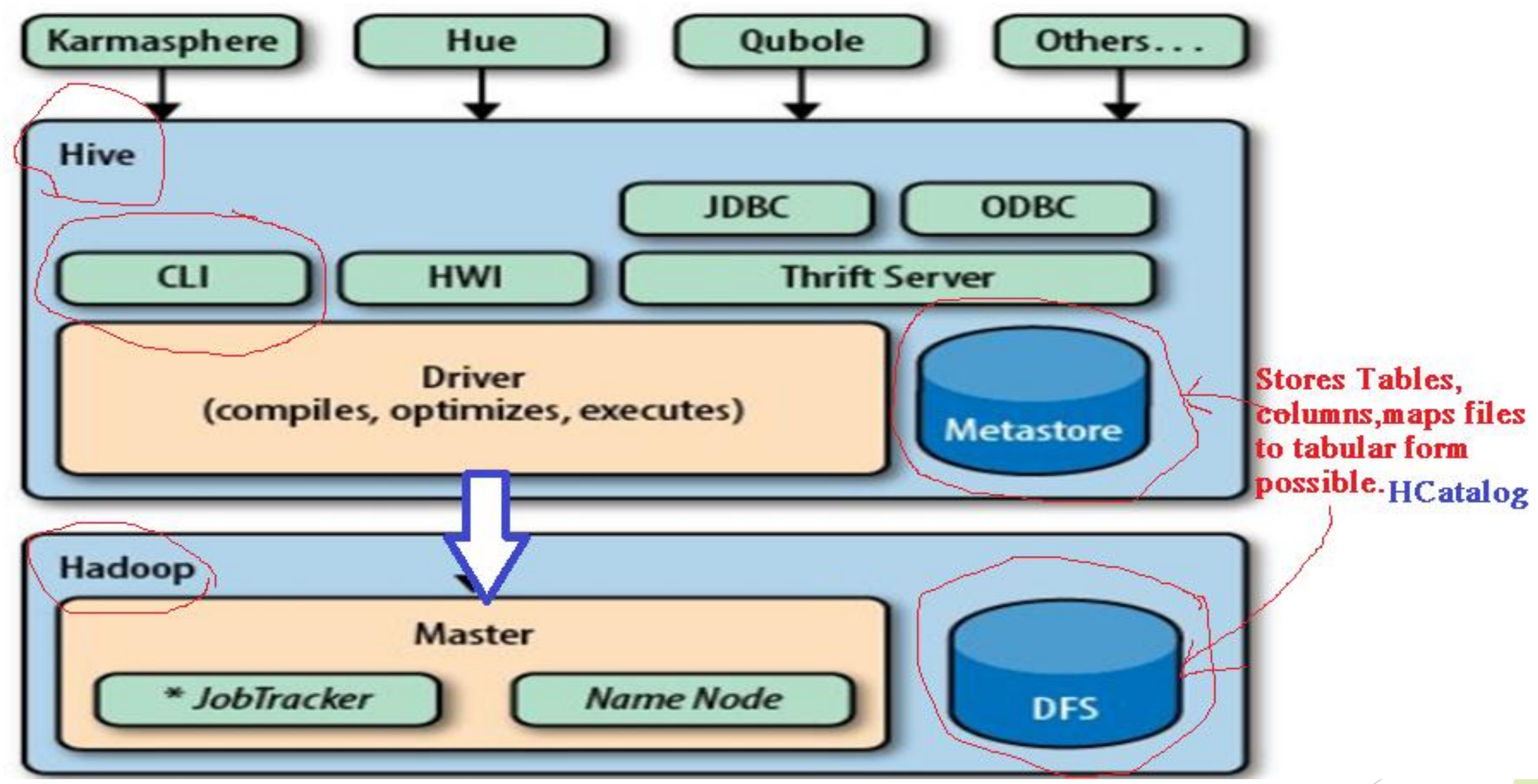
What can Hive do?

- Hive makes it easier to port SQL based application to Hadoop.
- Hive is most suited for data warehouse applications where relatively ***static*** data is analyzed.
- It does not provide OLTP(online transaction processing)
- Hive with Spark => **Shark => Spark SQL**

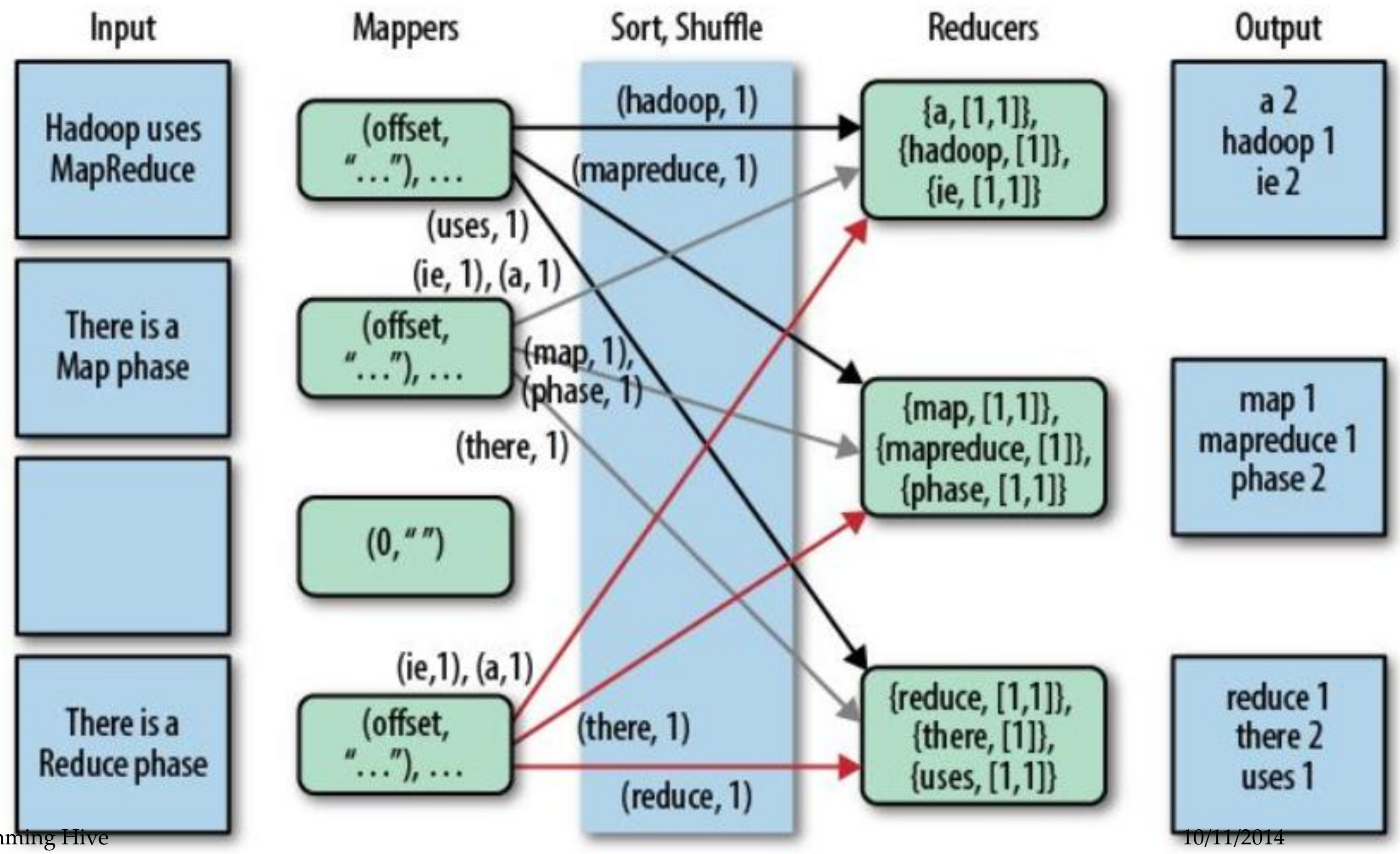
Hive vs RDMS

Hive	RDMS
For analytics and large aggregation	Real time processing
Data WareHouse	Database
High Latency, Fast load and flexibility	Fast Query
No transaction Support	Transaction support

Hive Components



Word Count Example Revisited



Word Count Example Code

- 63 lines of java code
- 25 lines of python code
- How many lines do you you think we will need to write the same in Hive programming Langauge?

Word Count Example Code

- 63 lines of java code
- 25 lines of python code
- How many lines do you you think we will need to write the same in Hive programming Langauge?
- 3 lines and 1 line with pre-proccssing!

Hive Directory Structure

- Lib Directory
  - $HIVE_HOME/lib
  - Location of Hive JAR file
  - Contains the actual Java code that implement the Hive functionality
- Bin Directory
  - $HIVE_HOME/bin
  - Location of Hive Scripts/Services
- Conf directory
  - HIVE_HOME/conf
  - Location of configuration files

[3] Accessing Hadoop data using Hive 10/11/2014 10

Hive Metastore

- What is Hive Metastore?
  - Hive metastore is a database that stores metadata about the hive tables. Metdata examples are *table name, column name, column type, table location* etc.
- Datastore
  - In process (Debry)
  - Out of process Datastores e.g DB2, MySQL, Oracle etc.
- Configuration for metastore:
  1. **Embedded**(same JVM)
  2. **Local** (out of process database)
  3. **Remote** (hive service JVM is separate from metastore JVM and database is remote place.

Hive Data units

- Database
- Table
- Partition
  - fundamentally horizontal slices of data which allow larges sets of data to be segmented into more manageable chunks.
- Buckets(Clusters)

/erp.db/customer/region=North America/country=US /erp.db/customer/region=North America/country=CA /erp.db/customer/region=South America/country=BR

Partition example

```
CREATE TABLE customer (
    id        INT, name        STRING, address1     STRING,address2     STRING,
    city        STRING,state        STRING, zip        STRING
)
PARTITION BY (
    REGION        STRING,
    country        STRING
);
```

Directory Structure in hdfs (for faster query)

/erp.db/customer/region=North America/country=US

/erp.db/customer/region=North America/country=CA

/erp.db/customer/region=South America/country=BR

Cluster and Bucket example

```
CREATE TABLE order (
    username     STRING,
    orderdate    STRING,
    amount       DOUBLE,
    tax        DOUBLE,
) PARTITIONED BY (company STRING)
CLUSTERED BY (username) INTO 25 BUCKETS;
```

- We are creating 25 buckets and clustering on ‘username’

Physical Layout

- Data files are regular HDFS files
- **Warehouse** Directory in HDFS Specified in hive-site.xml as **hive.metastore.warehouse.dir**

  e.g /home/hive/warehouse
- **Tables** stored in subdiretories of the warehouse
- **Partion** and **buckets** subdirectories of Table subdirectory.
- **Data** stored in flat file in **HDFS** with char delimited text or sequence file.

Hive DDL commands

- Create database

  CREATE DATABASE mydatabase;

  CREATE DATABASE mydatabase LOCATION '/myfolder/subdir/;

  CREATE DATABASE mydatabase COMMENT 'This is my database`;

- Delete database

  DROP DATABASE IF EXISTS mydatabase;

Data Types

- Primitive types
  - Integers
  - Boolean
  - String
  - Date/Time
  - Binary
- Complex Types
  - Arrays
  - Structs
  - Maps
  - Union

Hive Operators

- **Relational Operators**
  - SELECT id,name FROM users WHERE name **LIKE** ‘Tom%’;
    - Return values Tom ,Tomas

- **Logical Operators**
  - AND ,&&, OR, ||, NOT A , !A
  - Operators on complex Types
  - A[n] for array access
  - M[key] for map access
  - S.x for stuct field

Hive functions

- Built in Functions
  - count(*)
  - sum(col)
  - avg(col)
  - min(col),max(col)

SELECT sum(price) FROM fruits;
>27.75
SELEECT count(*) FROM furits;
> 4

Table "fruits"

apple	5.25
oranges	10.00
pipeapple	5.00
grapes	7.50

Comprehensive list of available functions
https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF

Code & demo

Hive Command line

- **CLI (Command line interface)**

```
>hive
 -d,--define <key=value> Variable substitution to apply to hive
                                         commands. e.g. -d A=B or --define A=B
 -e <quoted-query-string>          SQL from command line
 -f <filename>                              SQL from files
 -H,--help                                 Print help information
```

- **Beeline – New Command Line Shell**

  https://cwiki.apache.org/confluence/display/Hive/HiveServer2+Clients#HiveServer2Clients-Beeline–NewCommandLineShell

Hive Clients

- Command Line
- JDBC
- Python
- PHP
- ODBC

https://cwiki.apache.org/confluence/display/Hive/HiveClient

Hive FAQs

- **What is the default size of data processed per reducer in Hive?**
  - **1 G** controlled by using ***hive.exec.reducers.bytes.per.reducer***
- **How can I control number of mappers and reducers in Hive?**
  - ***mapred.max.split.size and mapred.max.split.size.***
  - ***hive.exec.reducers.bytes.per.reducer.***
- ***Check the reference below for more similar FAQS***
  - https://developer.ibm.com/hadoop/docs/getting-started/faqs/#hivfaq

Hive Alternatives

- **HBASE** – for rapid query and row level update ,transaction support
- Pig –High level data flow language
  - Higher learning curve for SQL programmers
  - Good for ETL , not good for ad-hoc query
  - Used in combination with Hive
- Java Map Reduce
  - Jave code for simple word count is about 63 lines.
  - More Hadoop internal architechture know how needed and quick prototyping difficult.

Other choices

- BigSQL ,BigR- from IBM
- Impala - from www.cloudera.com
- MapR- from www.mapr.com
- Cassandra - cassandra.apache.org

Big SQL

- **What is Big SQL?**
- Big SQL is a massively parallel processing (MPP) SQL engine that runs in Apache Hadoop to achieve vastly improved performance and SQL execution breadth over other SQL-on-Hadoop offerings.
- Big SQL 3.0 provides the following support over Hive 0.13:
  - More comprehensive SQL support (see below for details)
  - Federated queries
  - Statistics-driven optimization and query planning

Check the reference below for more detail feature comparison with Hive.

Which one to choose?

- Each has its advantage and disadvantages. You have to find out what is best for your application.
- Things to consider
  - Data Access Pattern
  - Volume
  - Cost
  - perfomence.
- Following is an article I found online
  - http://blog.markedup.com/2013/02/cassandra-hive-and-hadoop-how-we-picked-our-analytics-stack/

Hive Cheat Sheet for commands

- **Retrieving Information (General)**
  - SELECT from_columns FROM table WHERE conditions;
- **Retrieving All Values** SELECT * FROM table;
- **Retrieving Some Values** SELECT * FROM table WHERE rec_name = "value";
- **Retrieving With Multiple Criteria**
  - SELECT * FROM TABLE WHERE rec1 = "value1" AND rec2 = "value2";
- **Retrieving Specific Columns** SELECT column_name FROM table;
- **Retrieving Unique Output** SELECT DISTINCT column_name FROM table;
- **Sorting** SELECT col1, col2 FROM table ORDER BY col2;
- **Sorting Reverse** SELECT col1, col2 FROM table ORDER BY col2 DESC;
- **Counting Rows** SELECT COUNT(*) FROM table;
- **Grouping With Counting** SELECT owner, COUNT(*) FROM table GROUP BY owner;
- **Maximum Value** SELECT MAX(col_name) AS label FROM table;

References

1. **Programming Hive -**By Edward Capriolo, Dean Wampler, Jason Rutherglen (O'Reilly Media)
   - Edward Capriolo, Dean Wampler & Jason Rutherglen
2. **Bigdatauniversity :**
   1. **"Accessing Hadoop Data Using Hive"**
   2. **"SQL Access for Hadoop"**
3. **Apache Hive tutorial https://cwiki.apache.org/confluence/display/Hive/Tutorial**
4. **http://hortonworks.com/hadoop-tutorial/how-to-process-data-with-apache-hive/**
5. **https://developer.ibm.com/hadoop/docs/getting-started/faqs/**

THANK YOU!

“Big data is the new ***Natural resource****”* – Rometty, CEO IBM

ARE YOU READY to use this new resource?

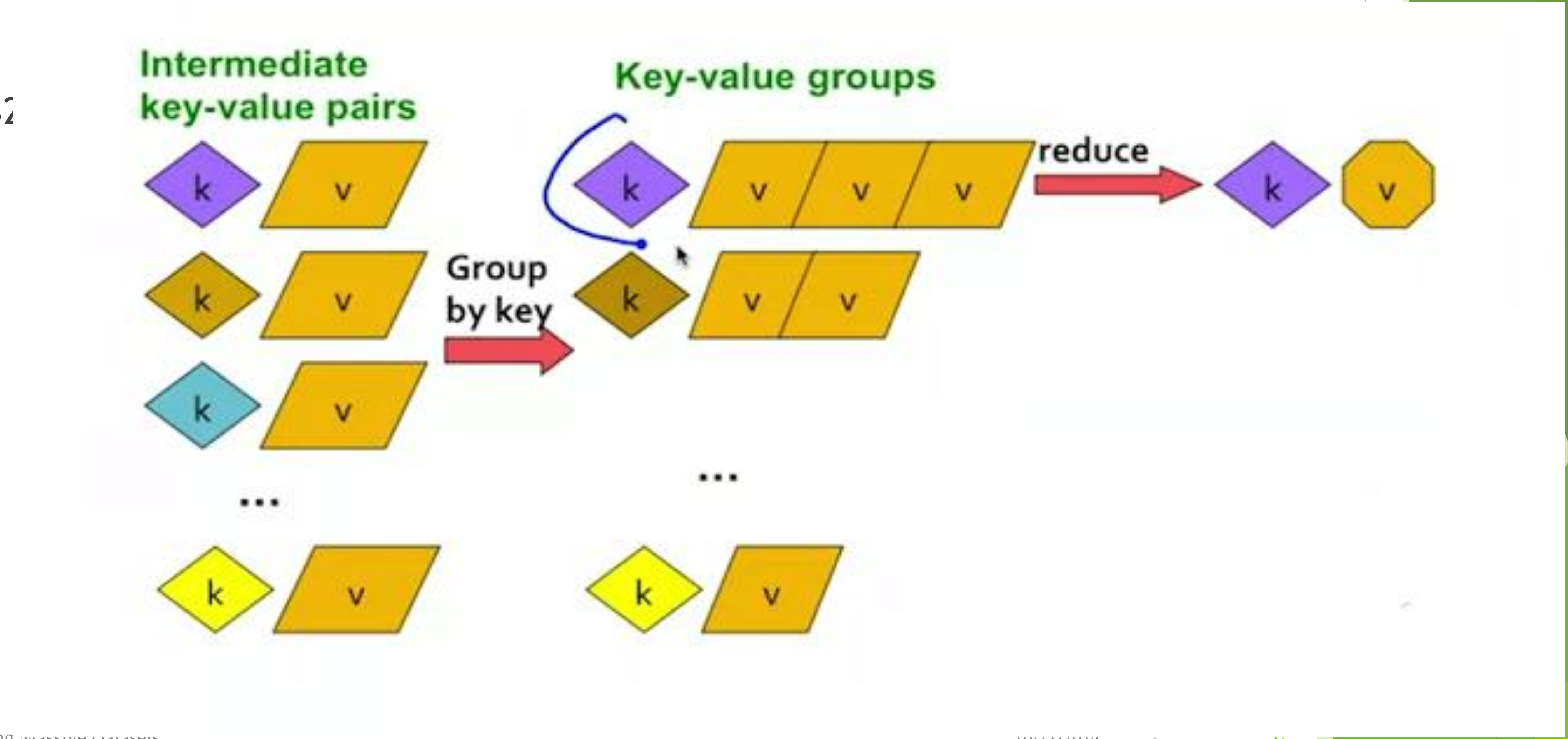
- Let me know if you want to volunteer for any presentation on any topic in future meetup.
- Any suggestion for future venue and topics?

Backup Slide

- Map – Reduce
  - Store data redundantly on multiple nodes for persistence and availability
  - Move computation close to data to minimize data movement
  - Simple programming model to hide the complexity of parallel computing in clusters of computer

Backup slide -Map Reduce Basics Revisited

- 32



Code for word Count

```
CREATE TABLE docs ( line STRING);

LOAD DATA INPATH ‘docs’ OVERWRITE INTO TABLE docs;

CREATE TABLE word_counts AS
SELECT word, count(1) AS count FROM
        ( SELECT explode(split(line, ‘\s’)) AS word FROM docs) w
        GROUP by word
        ORDER by word;
```