What it is Big Data

IDC Estimate
40 ZETABYTES
Of Data by 2020

2.5 QUINTILLION BYTES
of Data are Created each Day

200 BILLIONS PHOTOS

Global Healthcare data
150 EXABYTES
2.4 EXABYTES per year
Growth

Facebook has
1 PETTA BYTES
Of Storage

Hadron Collider generates
1 PETA BYTES
Of Data are create per year

VOLUME
Scale of Data

VARIETY
Different Forms of Data

Estimated
100 TERA BYTES
Of Data per US Company

Estimated
6 BILLION PEOPLE
Have a cell Phone

100 MILLION VIDEO
600 Years of Video
13 Hours of video uploaded per minute

500 MILLION TWEETS
Per day

BIG DATA

1.8 BILLION SMARTPHONES

Poor Data Quality cost businesses
600 BILLION A YEAR

1 IN 3 LEADERS
Don't trust the information they user to make decision

MAP

30% OF DATA COLLECTED
By marketers are not usable for real-time decision making

VELOCITY
Analysis of Data

VERACITY
Uncertainty of Data

20 BILLION NETWORK CONNECTIONS
By 2016

REDUCE

NY Stock Exchange generates
1 TERRA BYTES
Of Trade Data per day

Poor data across business and the government costs the US economy
3.1 TRILLION DOLLARS
a year

RESULT
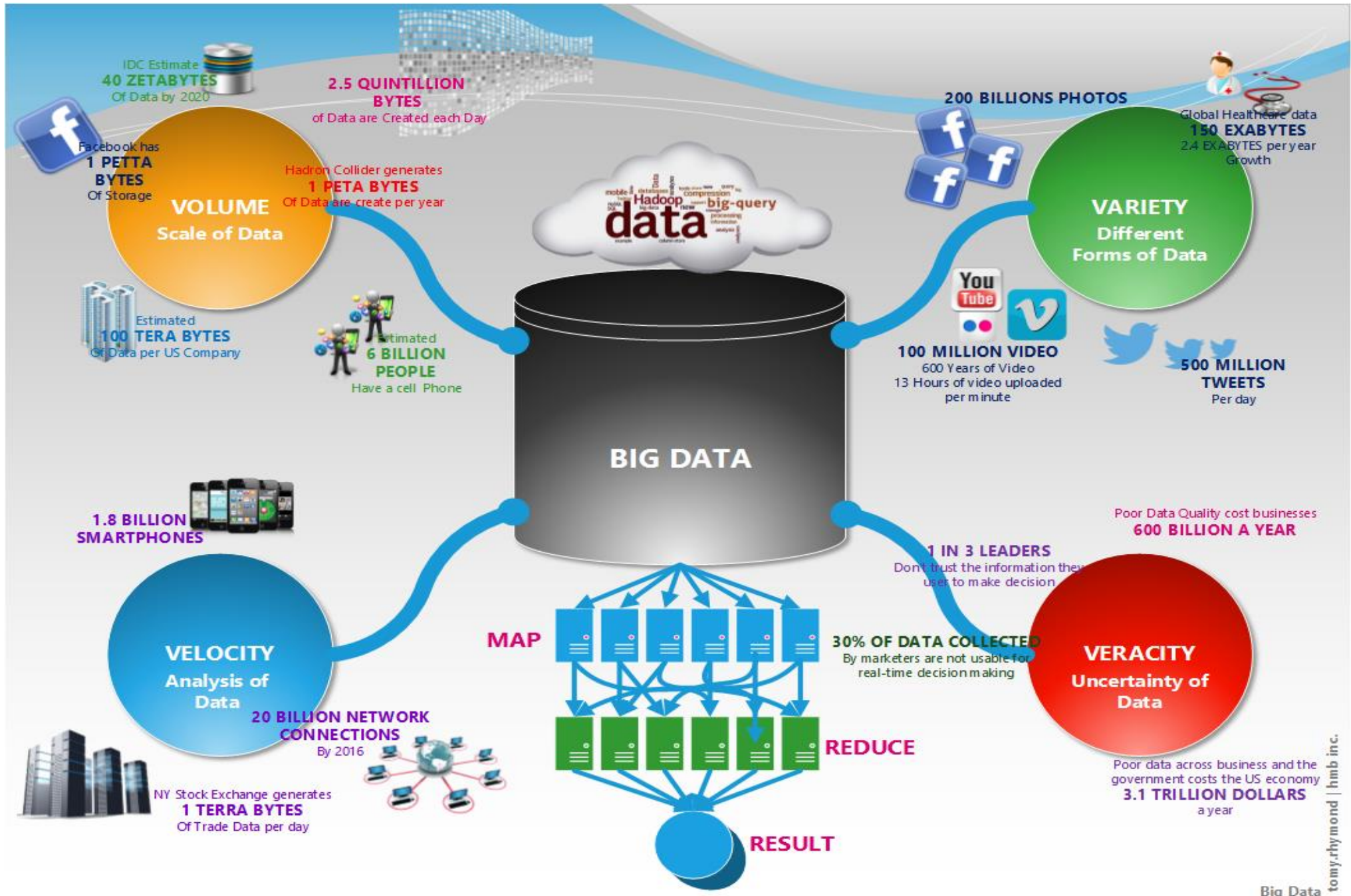
tomy.rhymond | hmb inc.

Big Data

WHY IS IT IMPORTANT NOW?

Smart Phones

4.6 billion mobile-phones.
1 - 2 billion people accessing the internet.

Connectivity:
Internet Of Things

Connectivity:
Social Networks

Facebook:1.06 bn monthly active users, 30 billion
pieces shared monthly.
~175 million tweets every day

The connectivity improved.
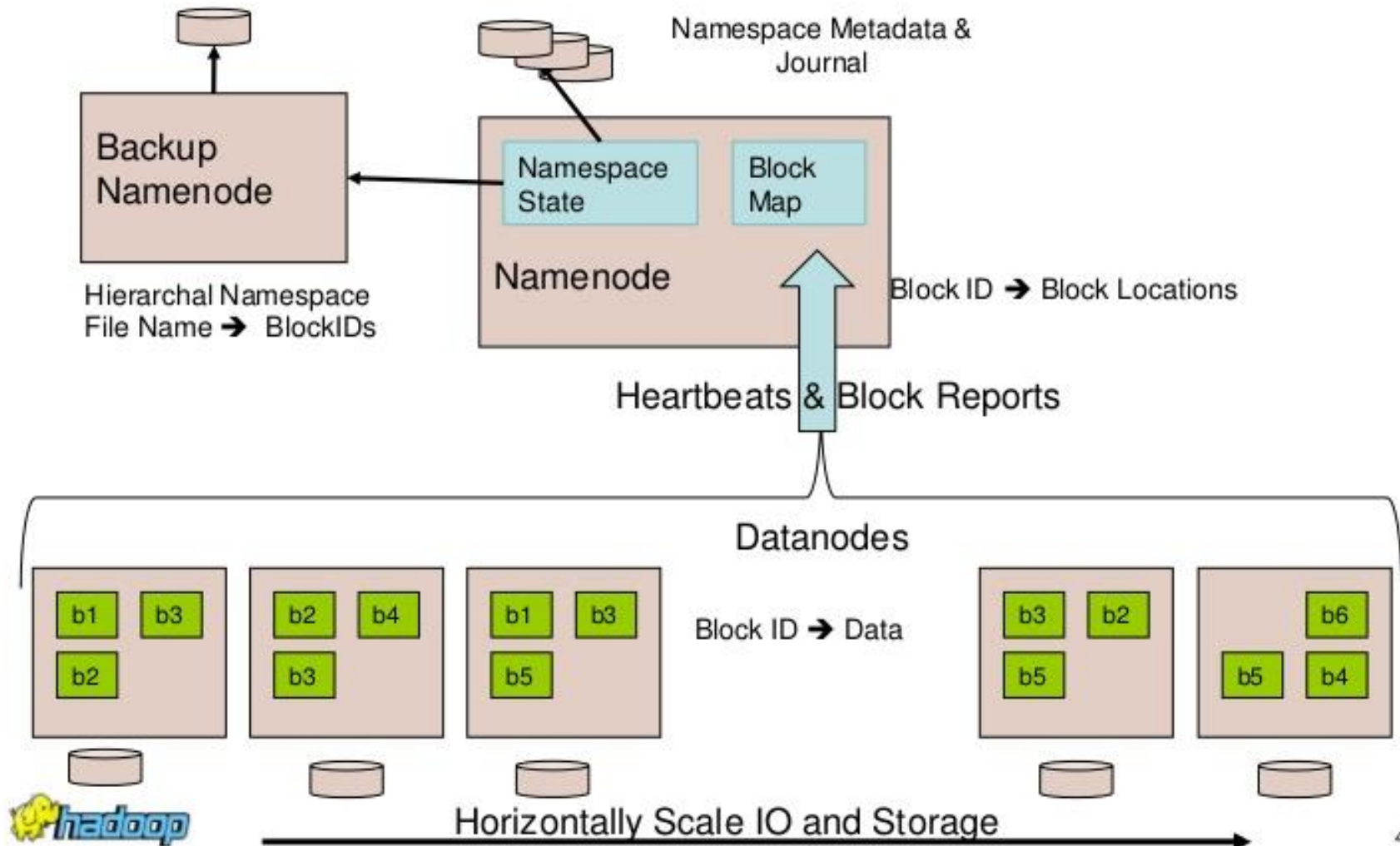The devices became cheaper, faster and smaller.
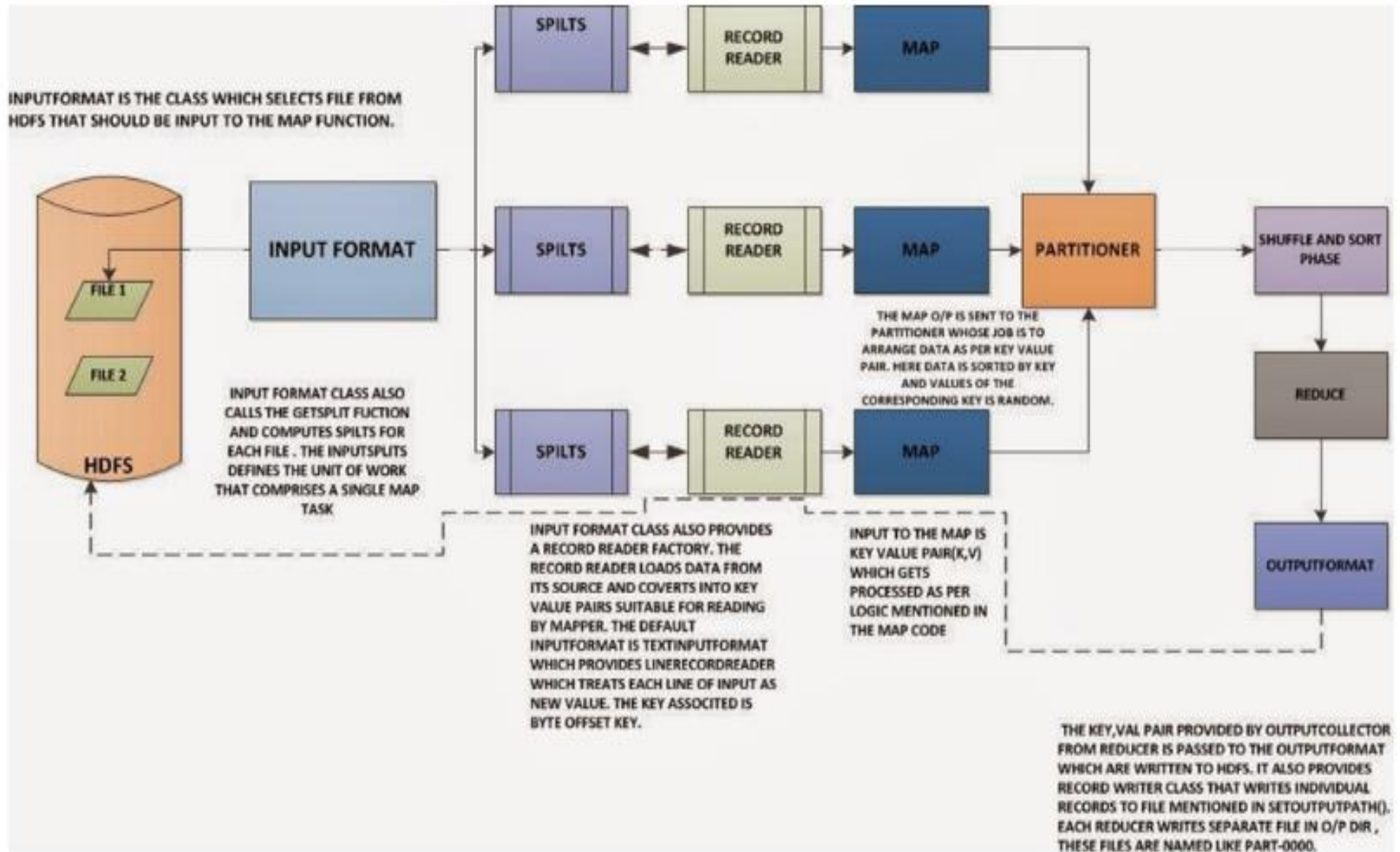
Hadoop
What is it and why does it matter?

- **Apache Hadoop** is an open-source software framework used for distributed storage and processing of very large data sets

- It consists of computer clusters built from commodity hardware

- All the modules in Hadoop are designed with a fundamental assumption that hardware failures are a common occurrence and should be automatically handled by the framework.

HDFS

Namespace Metadata & Journal

Backup Namenode

Namespace State

Block Map

Namenode

Hierarchal Namespace
File Name ➜ BlockIDs

Block ID ➜ Block Locations

Heartbeats & Block Reports

Datanodes

b1	b3	b2	b4	b1	b3
b2		b3		b5	

Block ID ➜ Data

b3	b2		b6
b5		b5	b4

Horizontally Scale IO and Storage

4

.Map Reduce Work Flow

INPUTFORMAT IS THE CLASS WHICH SELECTS FILE FROM HDFS THAT SHOULD BE INPUT TO THE MAP FUNCTION.

INPUT FORMAT CLASS ALSO CALLS THE GETSPLIT FUCTION AND COMPUTES SPILTS FOR EACH FILE . THE INPUTSPLITS DEFINES THE UNIT OF WORK THAT COMPRISES A SINGLE MAP TASK

THE MAP O/P IS SENT TO THE PARTITIONER WHOSE JOB IS TO ARRANGE DATA AS PER KEY VALUE PAIR. HERE DATA IS SORTED BY KEY AND VALUES OF THE CORRESPONDING KEY IS RANDOM.

INPUT FORMAT CLASS ALSO PROVIDES A RECORD READER FACTORY. THE RECORD READER LOADS DATA FROM ITS SOURCE AND COVERTS INTO KEY VALUE PAIRS SUITABLE FOR READING BY MAPPER. THE DEFAULT INPUTFORMAT IS TEXTINPUTFORMAT WHICH PROVIDES LINERECORDREADER WHICH TREATS EACH LINE OF INPUT AS NEW VALUE. THE KEY ASSOCITED IS BYTE OFFSET KEY.

INPUT TO THE MAP IS KEY VALUE PAIR(K,V) WHICH GETS PROCESSED AS PER LOGIC MENTIONED IN THE MAP CODE

THE KEY,VAL PAIR PROVIDED BY OUTPUTCOLLECTOR FROM REDUCER IS PASSED TO THE OUTPUTFORMAT WHICH ARE WRITTEN TO HDFS. IT ALSO PROVIDES RECORD WRITER CLASS THAT WRITES INDIVIDUAL RECORDS TO FILE MENTIONED IN SETOUTPUTPATH(). EACH REDUCER WRITES SEPARATE FILE IN O/P DIR , THESE FILES ARE NAMED LIKE PART-0000.