

Handling V's of Big Data

Velocity, Volume and Variety

Eka Kurniawan

BukaLapak

V's of Big Data

40 ZETTABYTES

[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE
have cell phones



WORLD POPULATION: 7 BILLION

Volume

SCALE OF DATA

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]
of data are created each day



Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session



By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

— almost 2.5 connections per person on earth



Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]



30 BILLION PIECES OF CONTENT
are shared on Facebook every month



Variety

DIFFERENT FORMS OF DATA



By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month



400 MILLION TWEETS
are sent per day by about 200 million monthly active users

**1 IN 3 BUSINESS LEADERS**

don't trust the information they use to make decisions



Poor data quality costs the US economy around
\$3.1 TRILLION A YEAR



27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

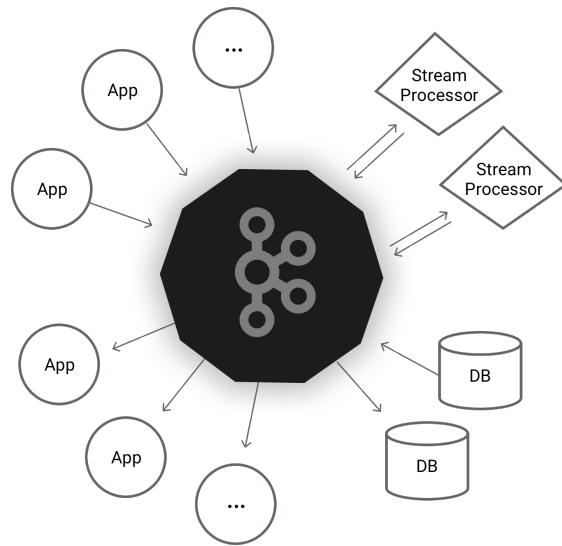
Veracity

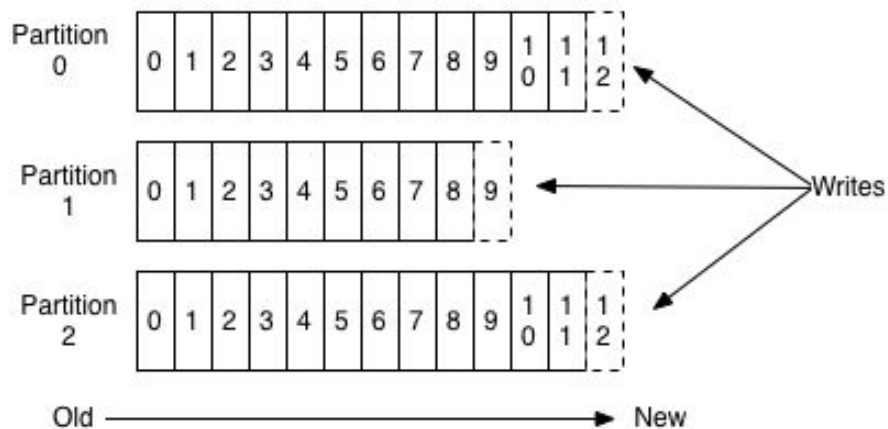
UNCERTAINTY OF DATA

Handling Velocity

Apache Kafka

distributed streaming platform

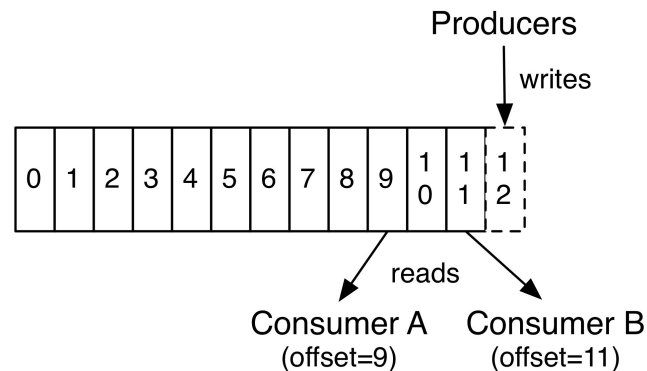




Anatomy of a Topic

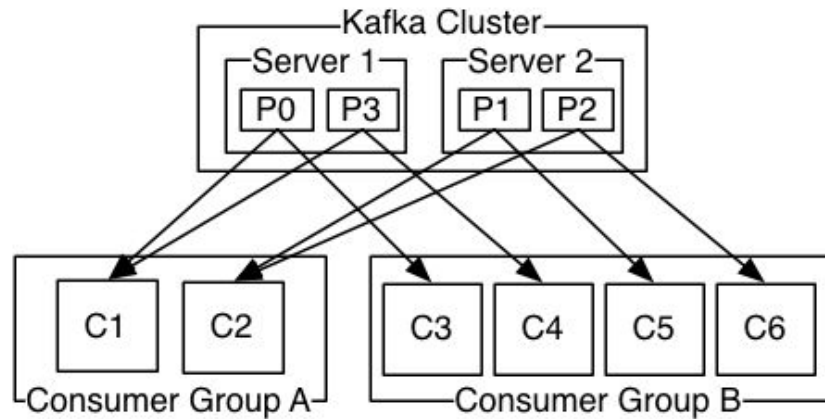
Topic

“category or feed name to which records are published”



Partition

“ordered, immutable sequence of records that is continually appended to”



Consumer Groups

Apache Kafka

Demo

Apache Kafka Benchmark Result

Produce and Consume Runtimes for Different
Number of Messages

Number of Messages	Total Messages Size (MB)	Produce Runtime (ms)	Consume Runtime (ms)
1	0.003	3	0.1
10	0.03	5	0.1
100	0.31	21	2
1000	3.13	173	20

Apache Kafka Benchmark Environment

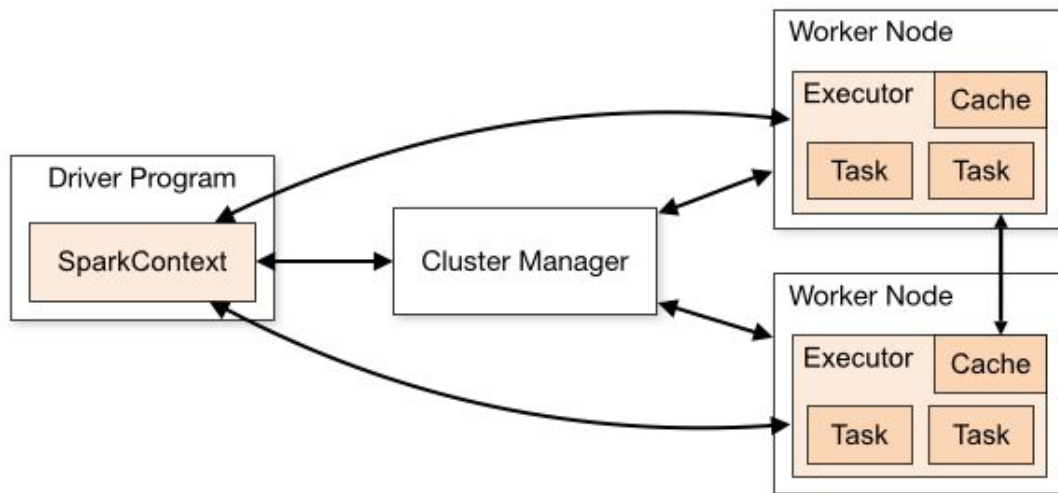
Hardware Specification and Software Version

Hardware	Specification	Software	Version
Processor	Intel(R) Core(TM) i7-3615QM CPU @ 2.30GHz	Python	3.6.4
CPU Cores	4	confluent_kafka	0.11.4
CPU Threads	8		
Memory	8 GB 1600 MHz DDR3		
Storage Device	SSD		

Handling Volume

Apache Spark

**unified analytics
engine for
large-scale data
processing**



Cluster Overview

SparkContext

“coordinates Spark applications run as independent sets of processes on a cluster”

Executors

“processes that run computations and store data for your application”

Apache Spark

Demo

Apache Spark Benchmark Result

Total Product Names to Process

Methods

Implementation

Total Product Names
per Thread per Second

Hashing

Python Dictionary

25

MapReduce

Apache Spark

191,167

Apache Spark Benchmark Environment

Hardware Specification and Software Version

Hardware	Specification	
	Hashing	MapReduce
Processor	Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz	Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz
Memory	256	16
Storage Device	HDD	HDD

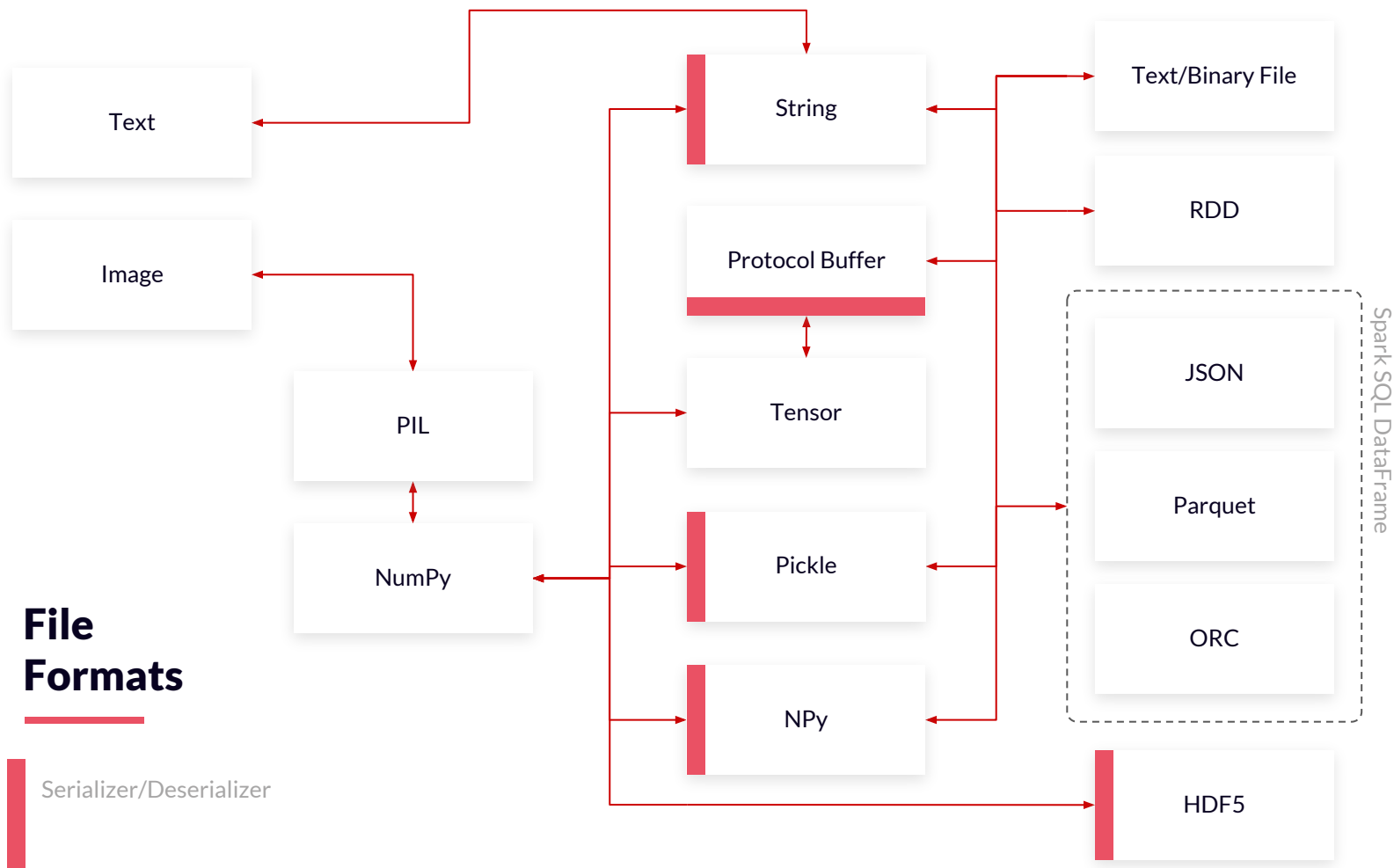
Software	Version	
	Hashing	MapReduce
Python	3.5.3	3.6.3
TensorFlow	1.9.0	1.11.0
PySpark	n/a	2.3.1

Handling Variety

Apache Spark SQL

**Apache Spark's
module for
working with
structured data**

File Formats



Serializers

Demo

Serializers Benchmark Result

Serializing and Deserializing 224x224 Matrix of
Integers

Serializers	Serializing Runtime (ms)	Deserializing Runtime (ms)	Storage Space (KB)
Protocol Buffer*	2.51	5.02	151
NPY*	0.787	0.557	269
Pickle*	0.103	0.076	151
HDF5**	3.78	1.87	153

* Write/read to/from memory

** Write/read to/from file

Serializers

Benchmark Environment

Hardware Specification and Software Version

Hardware	Specification
Processor	Intel(R) Core(TM) i7-3615QM CPU @ 2.30GHz
CPU Cores	4
CPU Threads	8
Memory	8 GB 1600 MHz DDR3
Storage Device	SSD

Software	Version
Python	3.6.6
NumPy	1.14.5
TensorFlow	1.10.1

File Formats

Demo

File Formats Benchmark Result

Image Preprocessing of 1000 same images with 4
executors, 4 cores/executor and Pickle as the
serializer

File Formats	Runtime/Image (ms)	Number of Partitions	Non-blocked Storage Space (MB)
JSON	22	20	809.4
Parquet	815	20	2.84
ORC	41	16	6.84

File Formats

Benchmark Environment

Hardware Specification and Software Version

Hardware	Specification	Software	Version
Processor	Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz	Python	3.6.3
		PySpark	2.3.0
		NumPy	1.14.0
		PIL	4.2.1

Conclusions

01 Big Data is here.

02 There are horizontally scalable solutions to handle the V's of Big Data.

03 Kafka produce can handle ~5 thousand messages/second.
Kafka consume can handle ~50 thousand messages/second.
Spark can handle ~191 thousand product names/thread/second.

04 No Silver Bullet. Data-mart stores convenient but redundant data.

05 Combination of Spark and TensorFlow is the foundation for Big Data and Big Compute.

Thank You



Eka Kurniawan

AI Engineer

