

BIG DATA ที่คุณคู่ควร เป็นแบบไหน
..จะใช้ **BIG DATA** แล้วต้องเริ่มอย่างไร ??
แล้วอีน ๆ อีกมากmany สำหรับความเข้าใจด้าน **BIG DATA**
มาบริหารความเข้าใจ และเริ่มก้าวแรกกับการใช้เทคโนโลยี **BIG DATA**



คับเพื่อป้อนใจกล้า
ท่องพากวนเรียนรู้แบบเป็นคันเอง

มองหลัง และหน้า **BIG DATA?** แล้วไปไหน ?



วันพฤหัสบดีที่ 4 มิถุนายน 58
เวลา 10.00-12.00 น.
ณ ห้องชมลุม

โดย คุณชัยวุฒิ สีทา
วศวกร หน่วยวิจัยสารสนเทศ การสื่อสารและการคำนวณ

สำรองที่นั่งล่วงหน้าได้ที่ งานประชาสัมพันธ์ email: pbrs@nnet.nectec.or.th
หรือลงทะเบียนได้บริเวณหน้างาน

ยินดีต้อนรับครับ

ชัยวุฒิ สีทา

วิศวกร
ICCRU
NECTEC

AGENDA

มองหลัง

เหลหน้า

BIGDATA

แล้วไปไหน

มองหลัง



นาย ชัยวุฒิ สีทา
COMPUTER SCIENCE
การทำงาน

- ออดีต Admin รพ.จังหวัดนครปฐม
- ปัจจุบันเป็นวิศวกรที่ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ประสบการณ์ทำงาน 9 ปี ด้านโปรแกรมมิ่ง และฐานข้อมูล สังกัด ICCRU LAB

ความเชี่ยวชาญ

1. การจัดการด้าน SERVER ADMIN ประเภท LINUX จัดการด้าน LAMP
2. การจัดการด้านฐานข้อมูล DATABASE ADMIN ด้าน MySQL , MariaDB, Postgresql ฐานข้อมูลแบบ NoSQL เช่น MongoDB , HBASE (HADOOP)
3. การจัดการและการออกแบบด้าน BIG DATA ด้วย CLOUDERA HADOOP
4. การพัฒนาโปรแกรมด้วยภาษา PHP , JAVA และอื่นๆ

ผลงานที่ภาคภูมิใจในอดีตและปัจจุบัน
ที่ได้ร่วมกับทีมพัฒนาและดูแลระบบ

จัดการด้าน Render Cluster และดูแลด้าน Website / Server

| Admin | | | | | | | | | |
|--|--|---|--|---|--|--|---|---|---|
| | | | | | | | | | |
| ท่าที่ ๑ ห้ามเดินเริ่วนั้น กล้ามเนื้อในหน้า ตา | ท่าที่ ๒ ประยุกต์มาราก ห้ามใช้ตัดตน แก้ลมข้อเมือ และ แก้ลมในล้ำสีรังค์ | ท่าที่ ๓ ประยุกต์มาราก ห้ามใช้ตัดตน แก้วดักท้องและ ข้อเท้า และแก้ ลมปวดศีรษะ | ท่าที่ ๔ ประยุกต์มาราก ห้ามใช้ตัดตน แก้วดักท้องและ ข้อเท้า และแก้เกียจ ลมปวดศีรษะ | ท่าที่ ๕ ประยุกต์มาราก ห้ามใช้ตัดตน แก้ลมเข็ปศีรษะและ ตามัว และแก้เกียจ ลมปวดศีรษะ | ท่าที่ ๖ ประยุกต์มาราก ห้ามใช้ตัดตน แก้ลมล่อน และ แก้เข้าข้อ | ท่าที่ ๗ ประยุกต์มาราก ห้ามใช้ตัดตน แก้ลมล่อนบัดคลาด และแก้เสื่อมหายเสบุก ร่างกาย | ท่าที่ ๘ ประยุกต์มาราก ห้ามใช้ตัดตน แก้วลมในแขน และแก้เสื่อมหายเสบุก ร่างกาย | ท่าที่ ๙ ประยุกต์มาราก ห้ามใช้ตัดตน ต่อจางหายอาชญา | ท่าที่ ๑๐ ประยุกต์มาราก ห้ามใช้ตัดตน แก้ให้ลี ชา และ แก้เข้า ชา |
| | | | | | | | | | |
| ท่าที่ ๑๑ ประยุกต์มาราก ห้ามใช้ตัดตน แก้โรคในอก | ท่าที่ ๑๒ ประยุกต์มาราก ห้ามใช้ตัดตน แก้ตะคริวเมือ ตะคริวเท้า | ท่าที่ ๑๓ ประยุกต์มาราก ห้ามใช้ตัดตน แก้ไข้หลี ตะโพกชั้ด และแก้ไข้หลี ตะโพกชั้ด | ท่าที่ ๑๔ ประยุกต์มาราก ห้ามใช้ตัดตน แก้ไข้ไข้หลี ตะโพกชั้ด และแก้ไข้หลี ตะโพกชั้ด | ท่าที่ ๑๕ ประยุกต์มาราก ห้ามใช้ตัดตน แก้ลมเดือนยันคามัว แก้เมือยันป่วยเมือ | ท่าที่ ๑๖ ประยุกต์มาราก ห้ามใช้ตัดตน ห้ามเดินรักษาตัว ปลายเท้า | ท่าที่ ๑๗ ประยุกต์มาราก ห้ามเดินรักษาตัว ปลายเท้า | ท่าที่ ๑๘ ประยุกต์มาราก ห้ามเดินรักษาตัว ปลายเท้า | ท่าที่ ๑๙ ประยุกต์มาราก ห้ามเดินรักษาตัว ปลายเท้า | ท่าที่ ๒๐ ประยุกต์มาราก ห้ามเดินรักษาตัว ปลายเท้า |

NECTEC
a member of NSTDA

ถ้าใช้ดีดตัน

ฉบับดิจิทัล

ความเป็นมา คำศัพท์ ความร่วมมือ

บริหาร/ร่วมพัฒนาระบบติดตามผู้สูญหายจากเหตุการณ์ต่าง ๆ
ในประเทศไทย 2554 - ปัจจุบัน Missingperson.or.th

เว็บเพจเชื่อมโยงระบบติดตามผู้สูญหายจากเหตุการณ์ต่าง ๆ ในประเทศไทย



Missing Children
ระบบติดตามเด็กหายออนไลน์ของประเทศไทย

ระบบติดตามผู้สูญหายออนไลน์จากเหตุอุทกภัยน้ำป่าในพื้นที่ภาคเหนือตอนบนกลาง

www.missingperson.or.th/flood

โครงการตามพระราชดำริสนับสนุนภารกิจช่วยเหลือผู้ประสบภัยทางอากาศฯ ขยายบูรณาการทุกมิติ



ระบบลงทะเบียนสัตว์สูญหาย
จากเหตุการณ์น้ำท่วมใหญ่ปี ๒๕๕๕



EMERGENCY.thai.net



สภากาชาดไทย
The Thai Red Cross Society

Registration for Missing Persons from Tsunami in Thailand

ระบบลงทะเบียนผู้สูญหาย จากเหตุการณ์สึนามิในประเทศไทย

This is an official website of Thailand for missing persons registration.

คลังข้อมูลสื่อประสม CORPUS ของ NECTEC

NECTEC
a member of NSTDA

สํานักงาน
นวัตกรรมแห่งชาติ

Annotated & Multimedia Corpus

ระบบคลังสื่อประสมและข้อความคำค้น

NECTEC
a member of NSTDA

ภาษาไทย ก ศ ท

You are here: Home

Main Menu

- บริการข้อมูล
 - ข้อมูลCorpus
 - คลังข้อความไทย-อังกฤษ
 - คลังพากไทย-อังกฤษ
 - คลังเสียงภาษาไทย
 - คลังข้อมูลภาษาพื้นเมือง
 - คลังปูร่างสามมิติคนไทย
 - คลังภาพจากกล้องจราจร
 - คลังฐานข้อมูลวินิจฉัย
 - คลังข้อมูล Anomaly Detection
- สำหรับสมาชิก

ค้นหา...

Login

ชื่อผู้ใช้งาน admin
รหัสผ่าน *****
จะจำชื่อและรหัสผ่าน
เข้าสู่ระบบ

ข้อมูลCorpus

คลังเสียงพูดไทย (Thai Speech Corpus)

R&D CORPUS

กพีชื่อความไทย-อังกฤษ (Thai-English Text Corpus)

กพีชื่อความไทย-อังกฤษ (Thai-English Text Corpus)

กพีเสียงพูดไทย (Thai Speech Corpus)

กพีชื่อความไทย-อังกฤษ (Thai-English Character Image Corpus)

กพีชื่อร่างกายมนุษย์ไทย (Thai 3D Body Model Corpus)

SizeThailand

การรักษาตนเป็น 3D Body

Dictionary

กพีภาษาไทย Lexicon

Thai OCR รักษาภาษาไทย

กพีชื่อบุคคลเจ้าของบัญชี Nucleus.CAM แบบ Trully รักษาบัญชี

Yaandyou.net Website / SERVER

NECTEC
a member of NSTDA

ສ່ວນະກິບ
NSTDA



The screenshot shows the Yaandyou.net website with a green header featuring the logo 'yo & you' and 'ยาดันคุณ'. The top navigation bar includes links for 'หน้าแรก', 'ร่วมรู้เรื่องยา', 'ร่วมรู้เรื่องโรค', 'ร่วมรู้สุขภาพ', 'ห้องนัดดิจิทัล', 'สำนักพัฒนา', 'แนะนำการเลือกค้น', and 'ผ่อนไถ่การไฟฟ้า'. Below the header is a search bar with placeholder text 'ເລືອນໄຊກາລັກນໍາທາງ : □ ນ້ຳມູນຄາ □ ນາກຄາວ' and a 'ຕັ້ງທາງ' button. A breadcrumb trail shows the user has navigated from 'หน้าแรก' to 'ยาดันคุณ' to 'ร่วมรู้เรื่องยา' to 'ยาดันคุณ' to 'ยาดันคุณ'. A sidebar on the left contains sections for 'แนะนำยาดันคุณ' and 'สุขภาพตามแพทย์และวัย', each with a list of links. The main content area features a QR code for the 'YaAndYou Application' available on the App Store and Google Play, along with download links for iPhone, Android, and Windows Phone. A right sidebar includes social media icons for Facebook, Twitter, and Email, a 'ผู้สนับสนุน' section, and a 'NECTEC' logo.

ภาพผลงานบางส่วน



ภาพผลงานบางส่วน



ระบบงานและฐานข้อมูลที่เคยจัดการ

- DAT ไฟล์หรือ TEXT ไฟล์ในการเขียนฐานข้อมูลกับ PASCAL / C พ.ศ. 2538 – 2539 File Handling

งานฐานข้อมูลกับ DOS PROGRAMMING

ชีวิตช่างໂหารร้าย !! โดยเฉพาะข้อมูลหลักล้านเรคอร์ด กับโครงการ เก็บข้อมูลการกินอาหารและอื่น ๆ ในฟาร์มไก่ ที่เลี้ยงในระบบปิด ...

ชีวิตเริ่มมีระดับขึ้น

- 2540 – 2544 ฐานข้อมูล / โปรแกรมพาก FOXPRO และ NETWARE ใน รพ.นครปฐม ข้อมูลหลักล้านเรคอร์ดเก็บยาวต่อเนื่องเป็นสิบปี

ปัญหาที่พบมากมาย การเขียน
โปรแกรมยังต้องใช้การปั่นรายงานเพื่อตอบ
โจทย์ KPI รายวันรายชั่วโมง

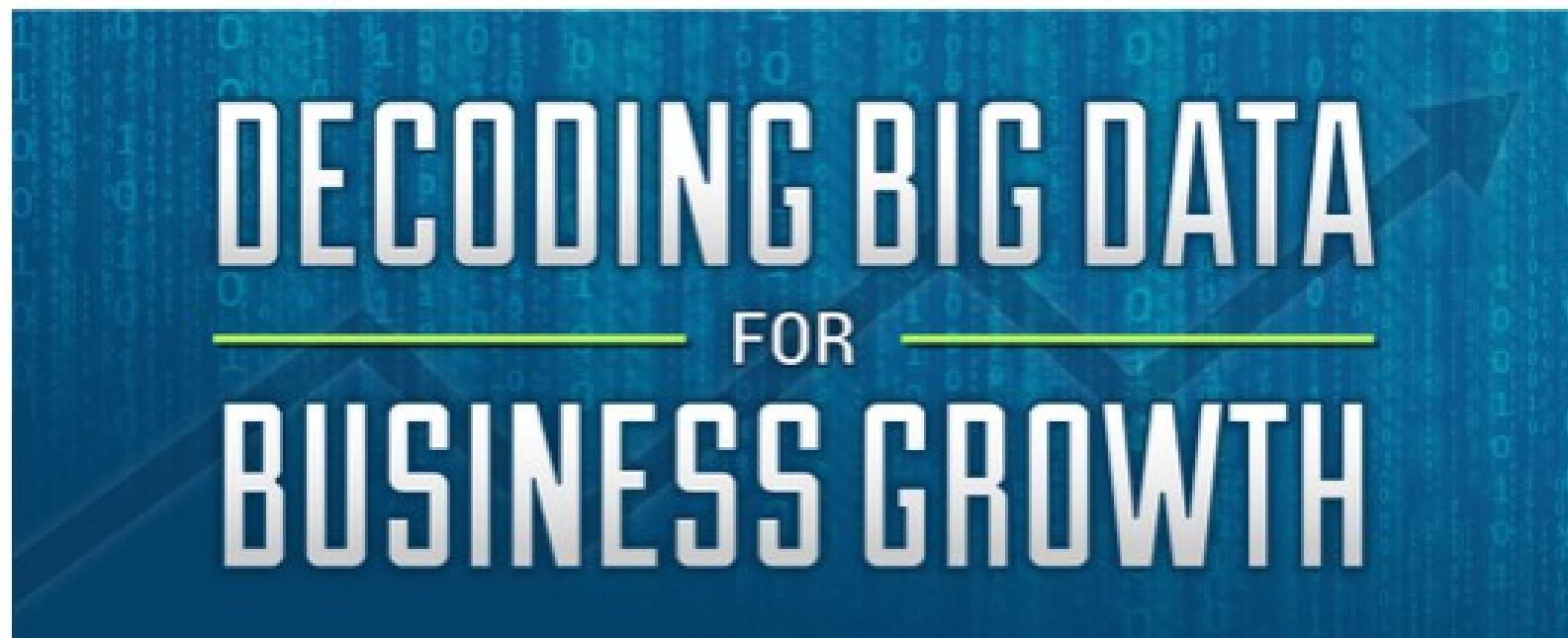
จะออกไปแตะขอบฟ้า

- 2548 ได้เข้าเมืองกรุงมาอยู่ NECTEC ได้รับงานด้าน MySQL และการจัดการเว็บไซต์ด้วย PHP / LINUX ADMIN การจัดการข้อมูลหลักล้านเรคอร์ดเริ่มง่าย ปั้นหาที่เคยเจอก็ได้ขึ้น โดยสรุปปัจจุบันมนุษย์งานพบร่ว่า ที่ผ่านมา
- เจอ กับ การ เก็บ ข้อมูล มาก มาก มาก กับ งาน ที่ พัฒนา
- ข้อมูล มี ทั้ง แบบ เป็น ข้อมูล ฐาน ข้อมูล หรือ ไม่ใช่ฐาน ข้อมูล
- มี TRANSACTION มาก มาก ใน ด้าน SERVER ที่ ต้อง เก็บ หรือ ดูแล

ยิ่งนานวันก็ยิ่งเป็นใหญ่เป็นโต
(งานและขนาดของงานที่เก็บนะ)
แล้วเราจะทำยังไงกันดี !!!

ແລກນ້າ

ค้นหาข้อมูลเกี่ยวกับ BIG DATA



Eric Siu
Contributor
CEO, Single Grain. Founder, Growth Everywhere.



COMPANIES COLLECT

>75,000

data points on a single individual

COMPANIES HAVE:



Invested in big data or plan to invest

64%



Take action on their collection of big data

48%



Use graphical images to represent data

45%



Design metrics and measures

42%

AND IT'S USED FOR:



Customer analytics



Experience analytics



Risk analysis



Regulatory compliance analysis



Location-based targeting



Campaign optimization

THE DATA COMES FROM:

Internal Data Sources



88%

Transactions



73%

Log Data



57%

Emails

External Data Sources



43%

Social Media



38%

Audio



34%

Photos & Video

ANALYTICS IS USED TO MAKE SENSE OF THE DATA AND IS CATEGORIZED IN FOUR WAYS



DESCRIPTIVE

Explains or describes what is happening

EXAMPLE

Demographic data such as gender, age, geography & income



DIAGNOSTIC

Explains or describes why something is happening

EXAMPLE

Divorce, birthday, weather, economy, etc.



PREDICTIVE

Explains or describes the probable outcome

EXAMPLE

How customers might respond to an ad



PRESCRIPTIVE

Explains or describes how to make something happen

EXAMPLE

Predicting gas prices so gas companies can lock in favorable terms

BIG DATA IN THE FORTUNE 1000

85% have big data initiatives already in place or are in planning

70% say their big data initiatives are focused on enterprise

75% say that multiple areas of business will be impacted

CLOUDERA BIG DATA CONCEPT



นี่คือภาพในฝัน และตอบโจทย์ได้อย่างเลอค่า สำหรับงานด้านข้อมูล

BIG DATA / Cloudera Administrator Training



ສເປົ່າຫໍ້ຂອງຜູ້ຄົນແລະເມືອງ ເມେລເບିର୍ବନ



อาหารการกินส์เต็ล์ไทย ๆ ใน เมลเบร์น



คอร์สการเรียนและสิ่งที่ได้รับ

HADOOP CLOUDERA ADMINISTRATOR

- ความเข้าใจในเรื่อง BIG DATA และมองว่าเราควรใช้มันหรือไม่ !!
- CLOUDERA ECOSYSTEM
- การออกแบบระบบ Cloudera HADOOP
- การติดตั้งและ Config ระบบ

BIG DATA គីឡូចារ៉ា ?

Big Data is a big thing. It will change our world completely and is not a passing fad that will go away. To understand the phenomenon that is big data, it is often described using **five Vs: Volume, Velocity, Variety, Veracity and Value**

เราควรใช้ BIG DATA หรือไม่ ?

มองงานเราคร่าวว่าใน 5V นี้ มี 3 ใน 5 หรือเปล่า คือ

1. VOLUME

ข้อมูลที่มีแนวโน้มในการเก็บข้อมูลในปริมาณมากและต่อเนื่อง เช่น ข้อมูลใน FACEBOOK, TWITTER , WEB SITE เป็นต้น (ข้อมูลเกิดเป็นรายวินาที, นาที เป็นต้น)

2. VELOCITY

ข้อมูลที่มีแนวโน้มในการเกิดข้อมูลในปริมาณมาก และต่อเนื่อง มีความสำคัญกับองค์กร

3. VARIETY

ข้อมูลที่มีความแตกต่างในด้านข้อมูล เช่น ข้อมูลฐานข้อมูล, ข้อมูลภาพ, ข้อมูลเดียง

4. VERACITY

ข้อมูลที่มีความน่าเชื่อถือ เกิดขึ้นแล้วสามารถนำมาใช้ผ่านกระบวนการสำคัญได้อย่างดี

5. VALUE

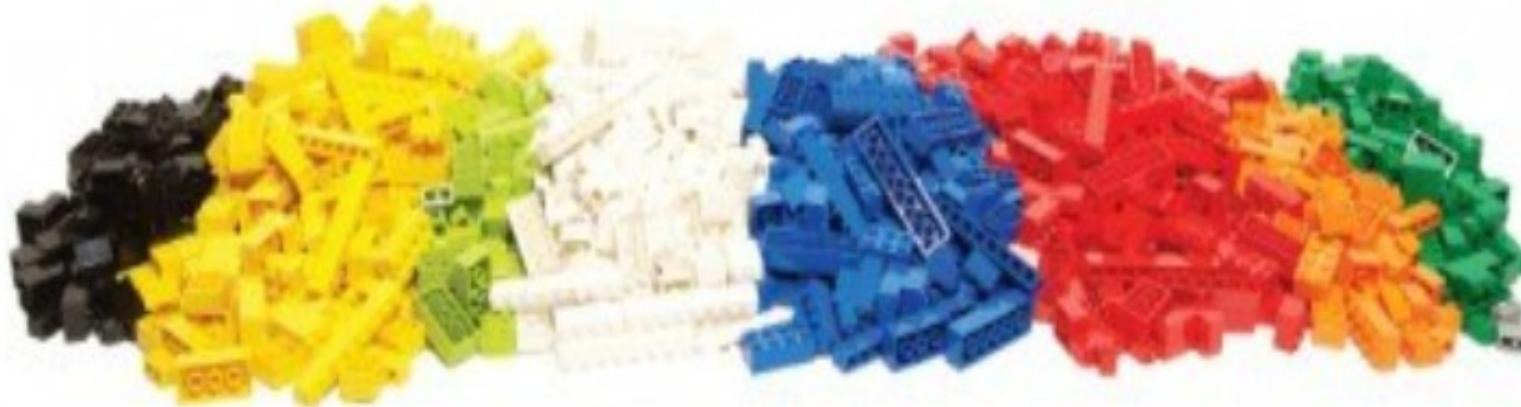
ข้อมูลที่มีคุณค่า มีมูลค่าต่อองค์กรและส่วนเกี่ยวข้อง

ความเข้าใจในเรื่อง BIG DATA



ถ้าข้อมูลคุณมีความหลากหลาย แต่มี 3 V นี้ คือ Volume Velocity และ Variety ให้สันนิษฐานก่อนว่า มันน่าจะเข้าข่าย BIG DATA ได้ และถ้าคุณวางแผนสร้าง ARCHIVE ด้วยลักษณะแบบแม่นแล้ว!!

SORTED



1. VOLUME

2. VELOCITY

3. VARIETY

ข้อมูลที่ Focus หลังจากเก็บเรียบร้อยในระบบแล้ว เราจะจัดเรียงข้อมูลด้วยการ Extract Transform LOAD (ETL) เข้าสู่ระบบฐานข้อมูลปกติ หรือเข้าสู่ฐานข้อมูลในรูปแบบ BIG DATA เช่น HBASE ได้ ซึ่งไฟล์ทั้งหลายจะถูกบรรจุในระบบไฟล์ของ HADOOP ที่เรียกว่า HDFS แต่ก็มีหลายที่ชอบที่จะเก็บ DATA ไปก่อนแล้วค่อยมา ETL ก็ยอมทำได้เช่นกัน

ARRANGED



1. VOLUME

2. VELOCITY

3. VARIETY

4. VERACITY

เมื่อจัดเรียงข้อมูลที่พร้อมใช้งานแล้ว ในขั้นนี้ข้อมูลจะเริ่มมีความน่าเชื่อถือในระดับการนำไปใช้งานได้ โดยกระบวนการนี้จะต้องกระทำภายใต้ HADOOP เช่น การเขียน Script Map Reduce เพื่อการจัดทำ ETL หรือคำนวณกระบวนการบางอย่าง ให้ได้รูปแบบที่จัดเรียงพร้อมใช้งาน

PRESENTED VISUALLY



1. VOLUME

2. VELOCITY

3. VARIETY

4. VERACITY

5. VALUE

คุณค่าที่คุ้มครอง เกิดจากการนำเสนอที่ควรคู่ ดังนั้น BIG DATA ที่มีคุณค่า ต้องเกิดจากการนำ DATA ที่เก็บดอง นำไปปรับรูปให้เกิดความเข้าใจของผู้อ่านข้อมูล นี่คือหัวใจที่สองของ BIG DATA ที่แยกมาจากหัวใจที่หนึ่ง นั่นคือการจัดเก็บข้อมูลขนาดใหญ่

The Awesome Ways Big Data Is Used Today To Change Our World

1. Understanding and Targeting Customers
2. Understanding and Optimizing Business Processes
3. Personal Quantification and Performance Optimization
4. Improving Healthcare and Public Health
5. Improving Sports Performance
6. Improving Science and Research
7. Optimizing Machine and Device Performance
8. Improving Security and Law Enforcement
9. Improving and Optimizing Cities and Countries
- 10. Financial Trading**

From Bernard Marr

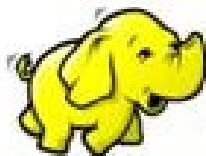
<https://www.linkedin.com/today/post/article/20131113065157-64875646-the-awesome-ways-big-data-is-used-today-to-change-our-world?trk=mp-author-card>

What Is Apache™ Hadoop®

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to **deliver high-availability**, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.





Apache Hadoop Ecosystem



Ambari

Provisioning, Managing and Monitoring Hadoop Clusters



Sqoop

Data Exchange



Zookeeper

Coordination



Oozie

Workflow



Pig

Scripting



Mahout

Machine Learning

R Connectors

Statistics



Hive

SQL Query



Hbase

Columnar Store

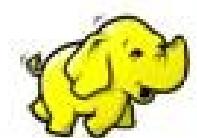


YARN Map Reduce v2

Distributed Processing Framework

HDFS

Hadoop Distributed File System



What is CLOUDERA ?

Cloudera offers a powerful and integrated Big Data platform comprising software, support, training, professional services, and indemnity. This platform, which has open source Apache Hadoop software at its core, allows customers to store, process, and analyze far more data, of more types and formats, and to do so more affordably than legacy technology -- allowing them to “ask bigger questions”.



CLOUDERA HADOOP

CDH

BATCH
PROCESSING
(MapReduce,
Hive, Pig)

ANALYTIC
SQL
(Impala)

SEARCH
ENGINE
(Cloudera Search)

MACHINE
LEARNING
(Spark, MapReduce,
Mahout)

STREAM
PROCESSING
(Spark)

3RD PARTY
APPS
(Partners)

WORKLOAD MANAGEMENT (YARN)

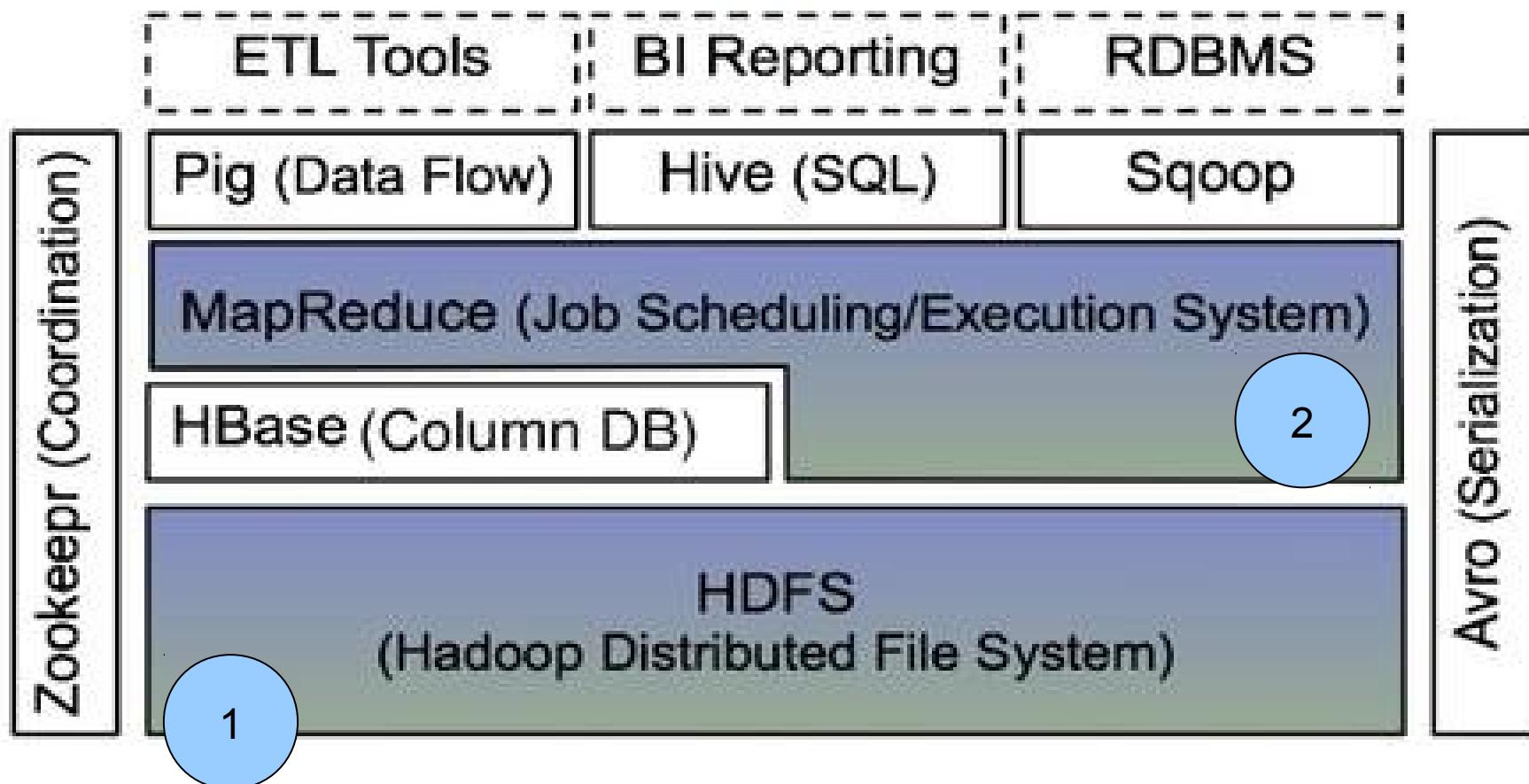
STORAGE FOR ANY TYPE OF DATA
UNIFIED, ELASTIC, RESILIENT, SECURE (Sentry)

Filesystem
(HDFS)

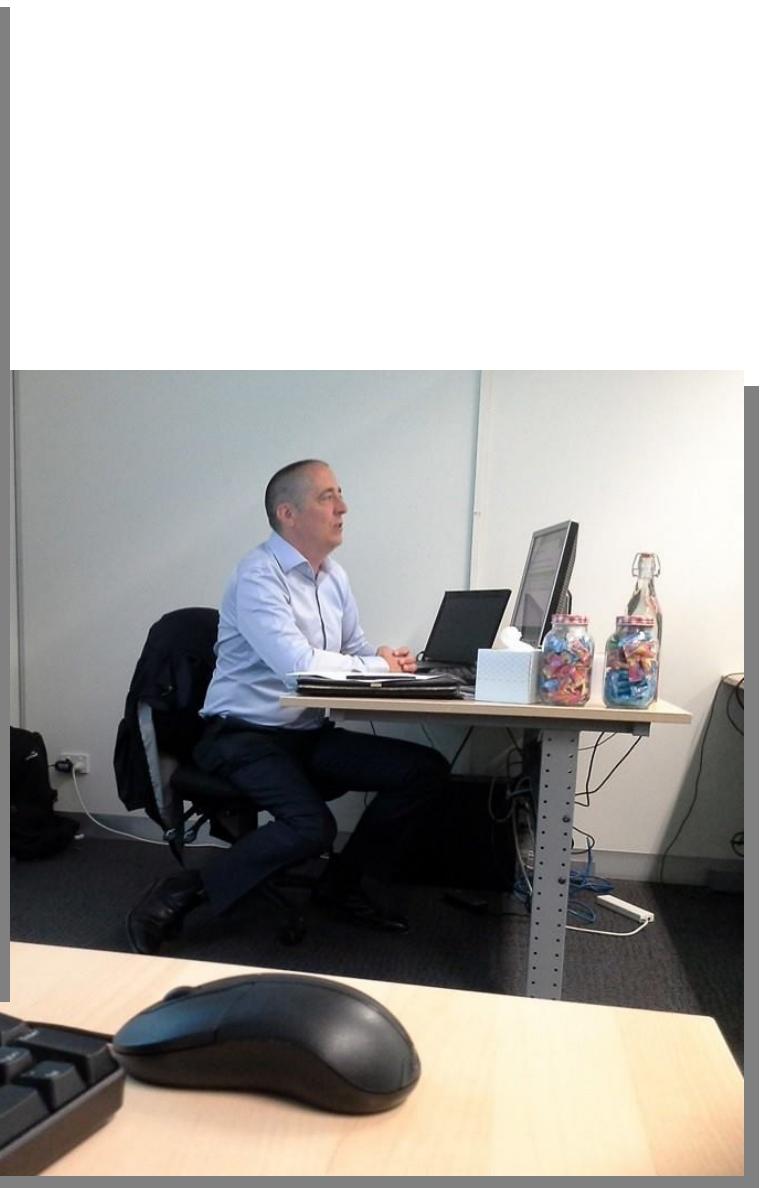
Online NoSQL
(HBase)

DATA INTEGRATION (Sqoop, Flume, NFS)

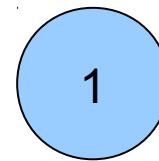
The Hadoop Ecosystem



Cloudera Administrator Training



Cloudera Administrator Training



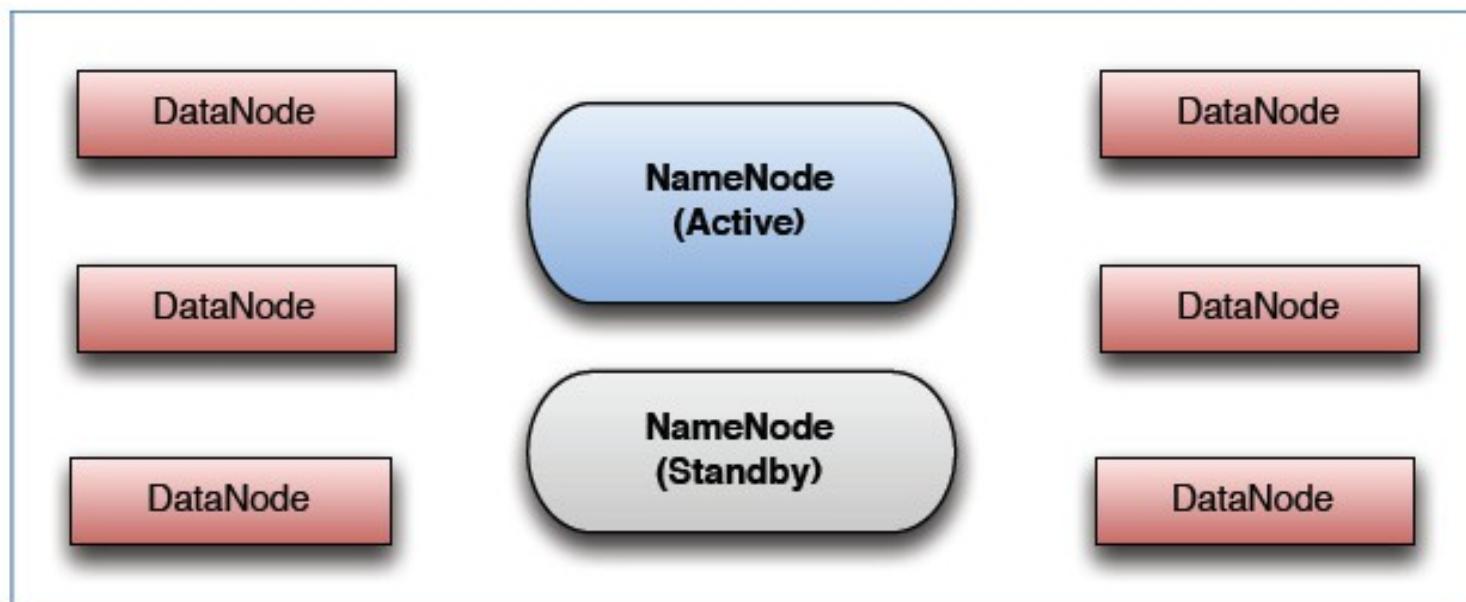
HDFS

HDFS Features

- **High performance**
- **Fault tolerance**
- **Relatively simple centralized management**
 - Master-slave architecture
- **Security**
 - Two levels from which to choose
- **Optimized for MapReduce processing**
 - Data locality
- **Scalability**

HDFS High Availability

- HDFS High Availability addresses the NameNode SPOF
- Two NameNodes: one active and one standby
 - Standby NameNode takes over when active NameNode fails
 - Standby NameNode also does checkpointing (Secondary NameNode no longer needed)



Cloudera Administrator Training

2

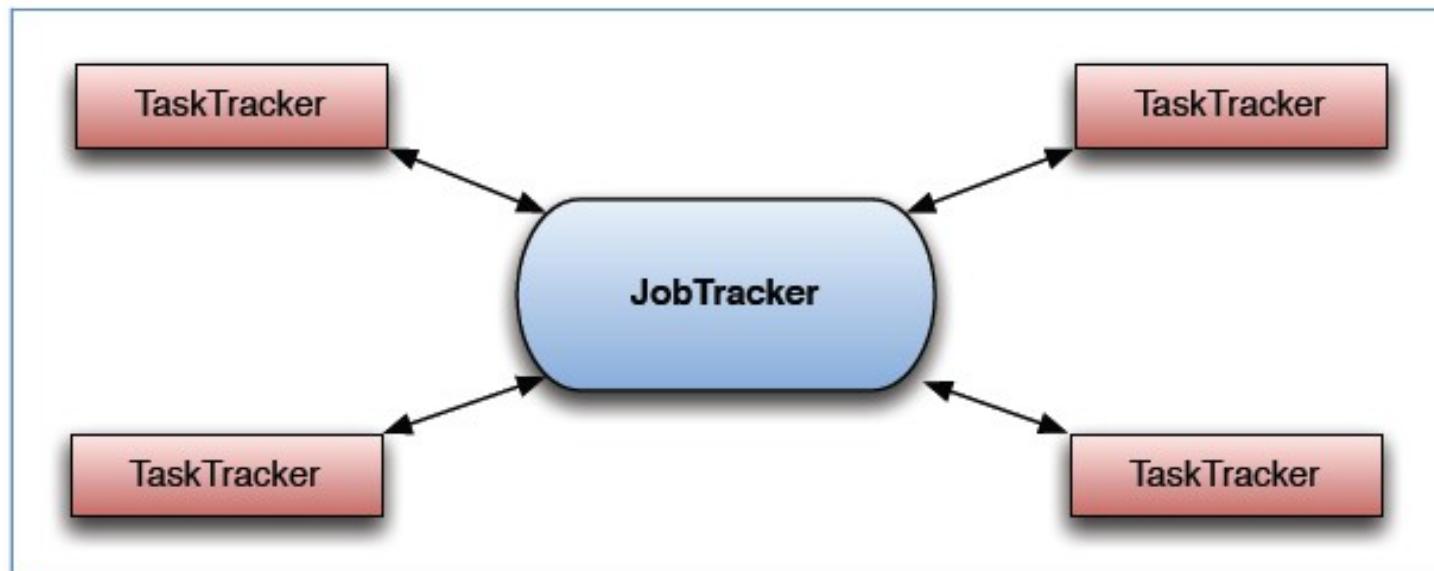
MapReduce

What Is MapReduce?

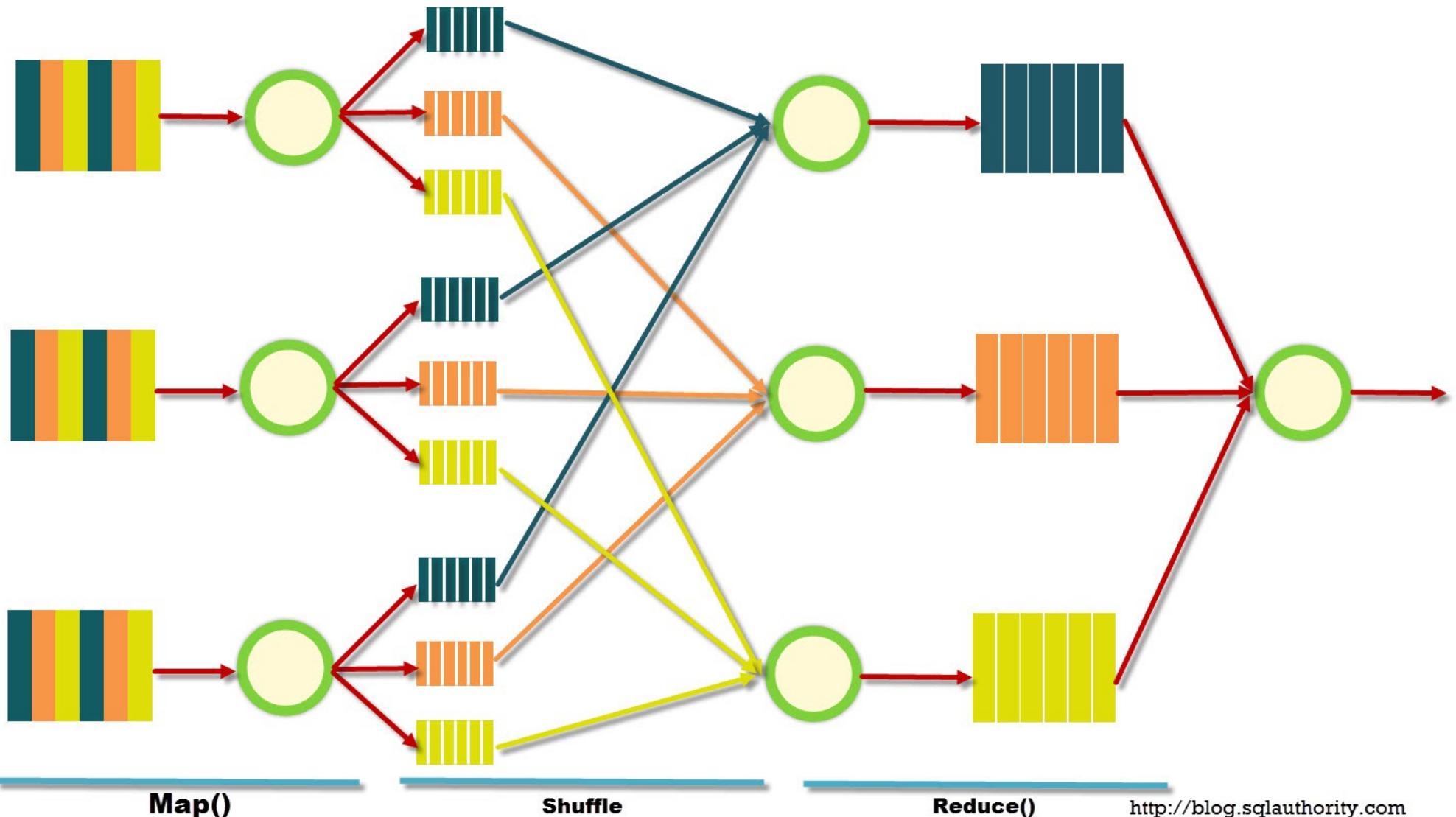
- **MapReduce is a programming model**
 - Neither platform- nor language-specific
 - Record-oriented data processing (key and value)
 - Facilitates task distribution across multiple nodes
- **Where possible, each node processes data stored on that node**
- **Consists of two developer-created phases**
 - Map
 - Reduce
- **In between Map and Reduce is the *shuffle and sort***
 - Sends data from the Mappers to the Reducers

Architectural Overview

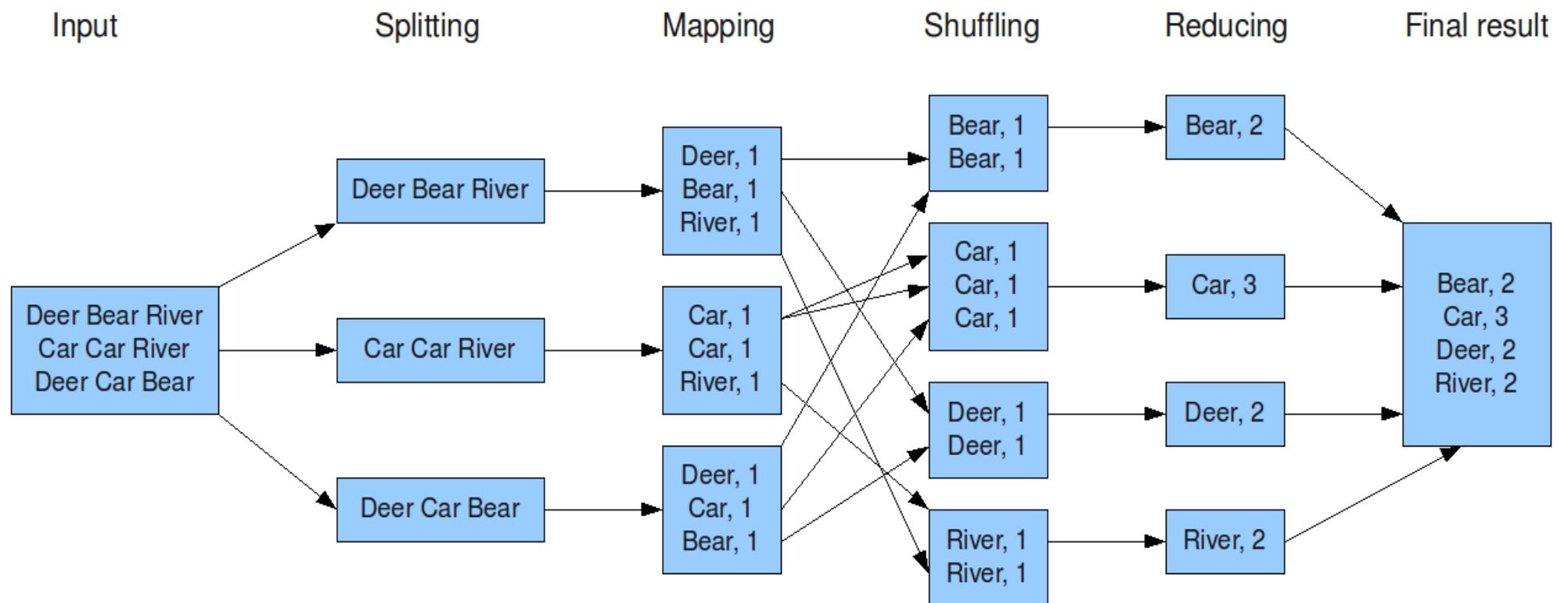
- There are two daemons in “classical” MapReduce
 - JobTracker (master) - exactly one per cluster
 - TaskTracker (slave) - one or more per cluster
- Slave nodes run both a TaskTracker and a DataNode daemon



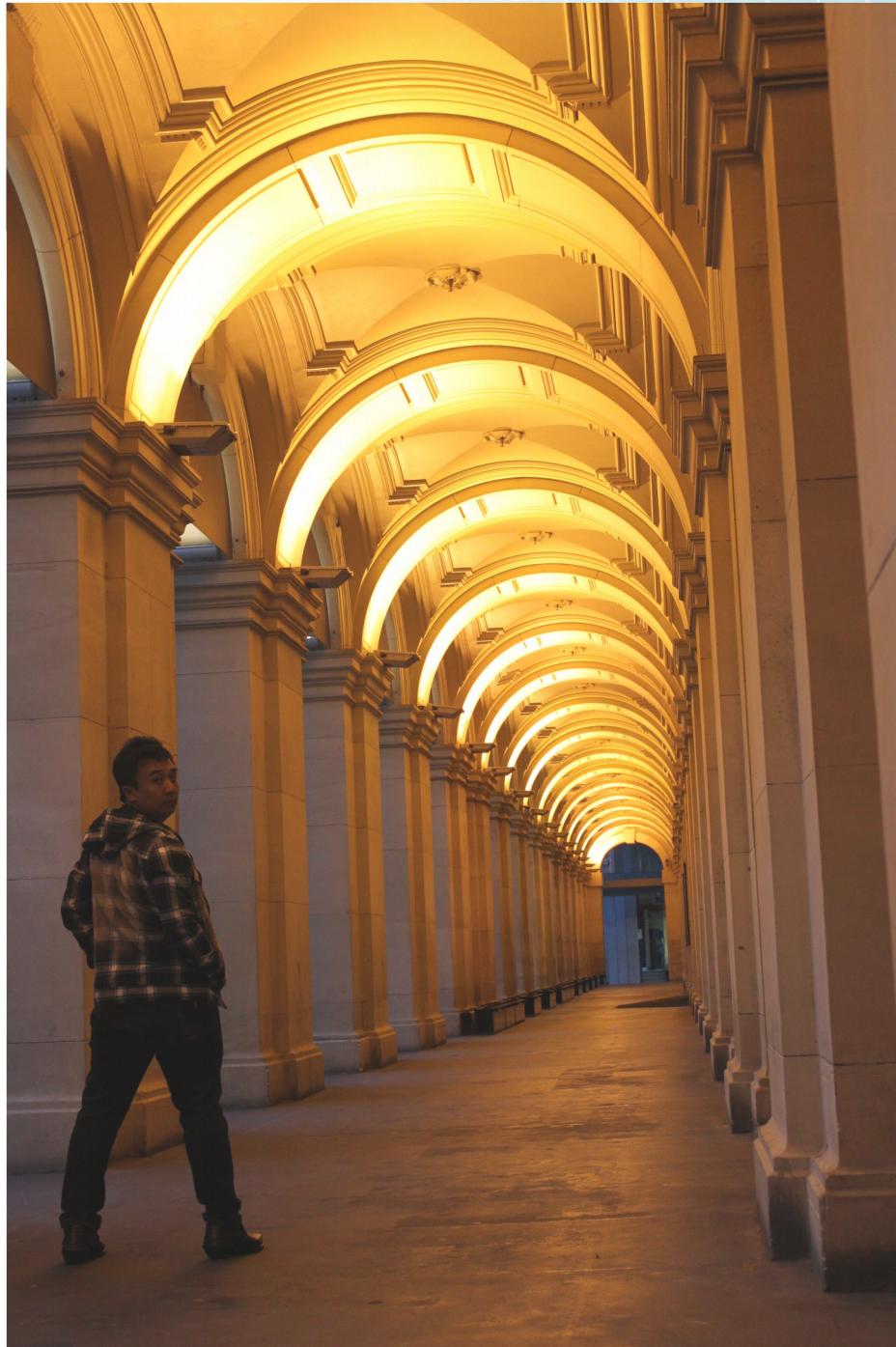
How MapReduce Works?



The overall MapReduce word count process



สิ่งที่ได้รับจากการเทรนนิ่งครั้งแรก



- สามารถติดตั้งระบบ HADOOP ของ Cloudera แบบ MANUAL ได้
- สามารถออกแบบสถาปัตยกรรมของการ ว่างระบบ Hadoop ได้
- สามารถติดตั้ง Service ต่างๆ ของ Hadoop ได้แบบ Manual
- มีเพื่อนจาก Cloudera Support ที่คอย ช่วยเหลือและแนะนำเทคนิคดี ๆ เป็น Shortcut ที่ดีให้กับชีวิต
- อาหารไทย อร่อยที่สุดในโลก



Cloudera Data Analyst Training: Using Pig, Hive, and Impala with Hadoop

16-19 FEB 2015 @ SYDNEY







Certificate of Attendance

is hereby granted to

chaiwoot seetha

To verify that he/she has attended

**Cloudera Data Analyst Training:
Using Pig, Hive, and Impala with Hadoop**

A handwritten signature in black ink that reads "Sarah Spiechale".

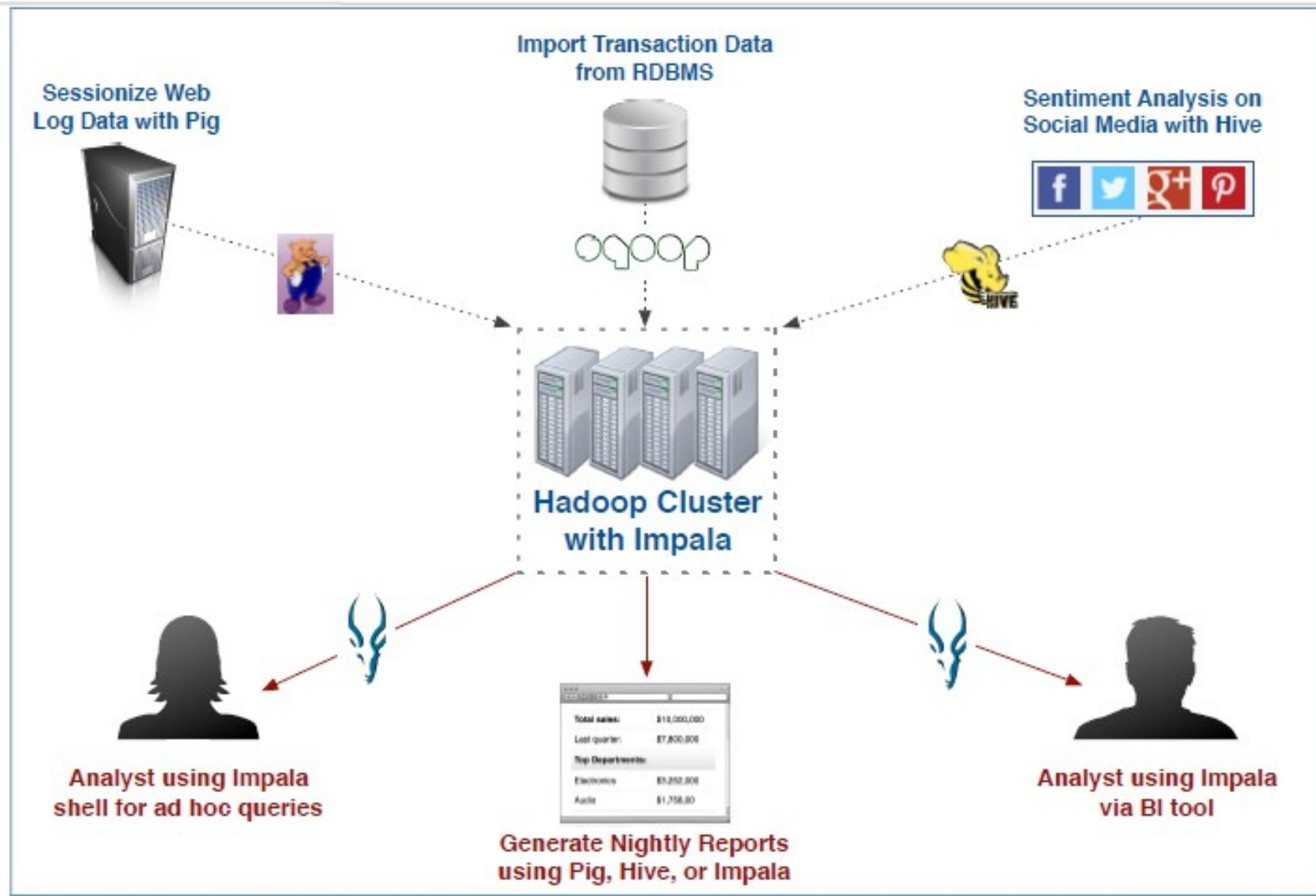
VP, Educational Services

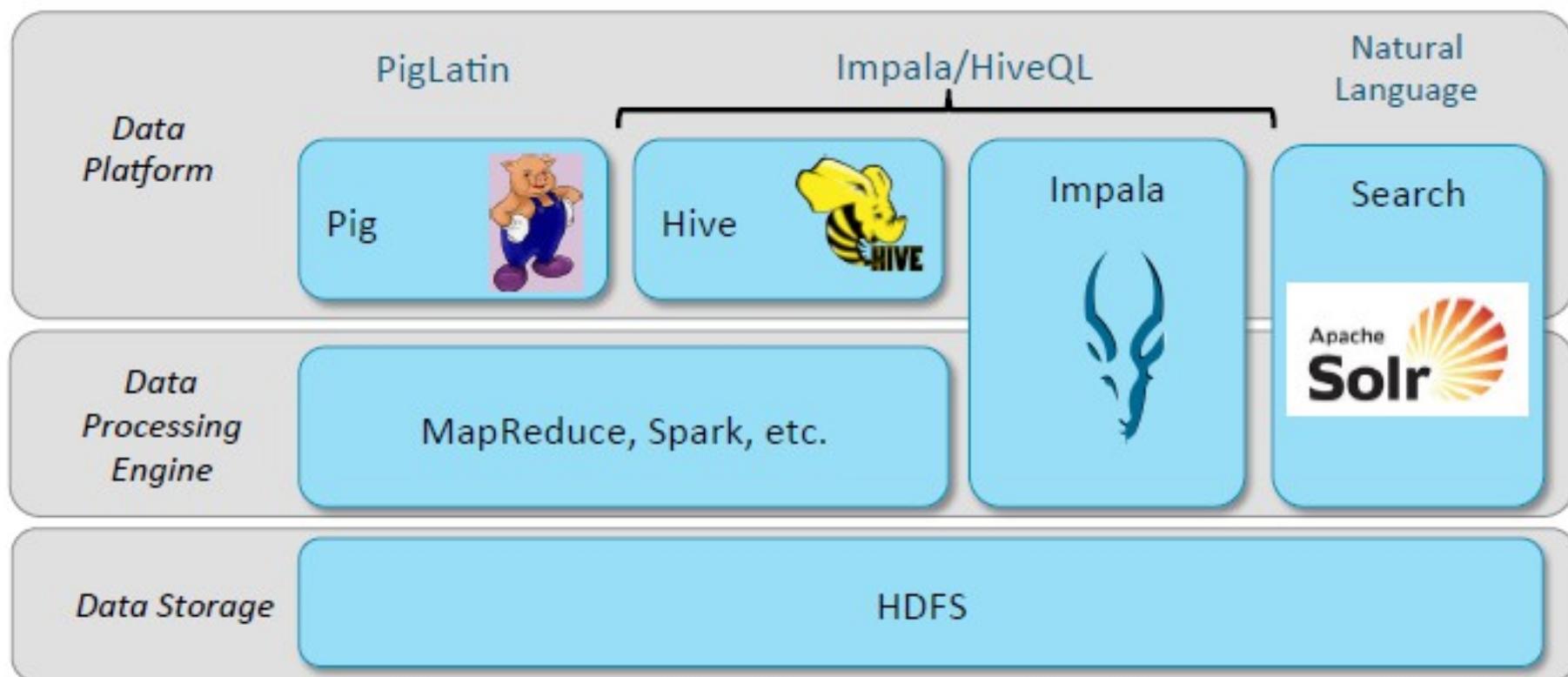
Cloudera, Inc.
www.cloudera.com

February 16th 2015

Course Date

Analysis Workflow Example





Apache Pig

- Apache Pig builds on Hadoop to offer high-level data processing
 - This is an alternative to writing low-level MapReduce code
 - Pig is especially good at joining and transforming data

```
people = LOAD '/user/training/customers' AS (cust_id, name);
orders = LOAD '/user/training/orders' AS (ord_id, cust_id, cost);
groups = GROUP orders BY cust_id;
totals = FOREACH groups GENERATE group, SUM(orders.cost) AS t;
result = JOIN totals BY group, people BY cust_id;
DUMP result;
```

- The Pig interpreter runs on the client machine
 - Turns PigLatin scripts into MapReduce jobs
 - Submits those jobs to the cluster



Apache Hive

- Hive is another abstraction on top of Hadoop
 - Like Pig, it also reduces development time
 - Hive uses a SQL-like language called HiveQL

```
SELECT customers.cust_id, SUM(cost) AS total
      FROM customers
      JOIN orders
        ON (customers.cust_id = orders.cust_id)
 GROUP BY customers.cust_id
 ORDER BY total DESC;
```

- A Hive Server runs on a master node
 - Turns HiveQL queries into MapReduce jobs
 - Submits those jobs to the cluster



Cloudera Impala

- Massively parallel SQL engine which runs on a Hadoop cluster
 - Inspired by Google's Dremel project
 - Can query data stored in HDFS or HBase tables
- Uses Impala SQL
 - Very similar to HiveQL
- High performance
 - Typically at least 10 times faster than Hive or MapReduce
 - High-level query language (subset of SQL-92)
- Impala is 100% Apache-licensed open source



Comparing Pig, Hive, and Impala

| Feature | Pig | Hive | Impala |
|------------------------------------|------|------|--------|
| SQL-based query language | No | Yes | Yes |
| Optional schema and metastore | Yes | No | No |
| User-defined functions (UDFs) | Yes | Yes | Yes |
| Process data with external scripts | Yes | Yes | No |
| Extensible file format support | Yes | Yes | No |
| Complex data types | Yes | Yes | No |
| Query latency | High | High | Low |
| Built-in data partitioning | No | Yes | Yes |
| Accessible via ODBC / JDBC | No | Yes | Yes |

Comparing an RDBMS to Hive and Impala

| Feature | RDBMS | Hive | Impala |
|---------------------------|-----------|--------------|--------------|
| Insert individual records | Yes | No | Yes |
| Update and delete records | Yes | No | No |
| Transactions | Yes | No | No |
| Role-based authorization | Yes | Yes (Sentry) | Yes (Sentry) |
| Stored procedures | Yes | No | No |
| Index support | Extensive | Limited | None |
| Latency | Very low | High | Low |
| Data size | Terabytes | Petabytes | Petabytes |
| Complex data types | No | Yes | No |
| Storage cost | Very high | Very low | Very low |

การทดสอบประสิทธิภาพ HADOOP

Impala Vs Hive Performance Testing

WITH

AMAZON EMR DATA TEST

7 May 2015

Chaiwoot Seetha

ICCRU NECTEC

ขั้นตอนการทดสอบ

1. จัดทำฐานข้อมูลทดสอบโดยโปรแกรมจาก AMAZON
โดยมีข้อมูล 3 TABLE มีขนาดของไฟล์ข้อมูลและ ROW ดังนี้

| Input Class (size of each table) | Books table (Million Rows) | Customers table (Million Rows) | Transactions table (Million Rows) |
|-------------------------------------|-------------------------------|-----------------------------------|--------------------------------------|
| 4 GB | 63 | 53 | 87 |
| 32 GB | 497 | 419 | 659 |

Query และตรวจสอบ Execution time

3. สรุปผลเปรียบเทียบ

Q1: Scan Query

```
SELECT COUNT(*)  
      FROM Table name  
      WHERE Field = Value;
```

This query performs a table scan through the entire table.

With this query, we mainly test:

- * Impala's read throughput compared to that of Hive.
- * With a given aggregated memory size, is there a limit on input size when performing a table scan, and if yes, what is the maximum input size that Impala can handle?

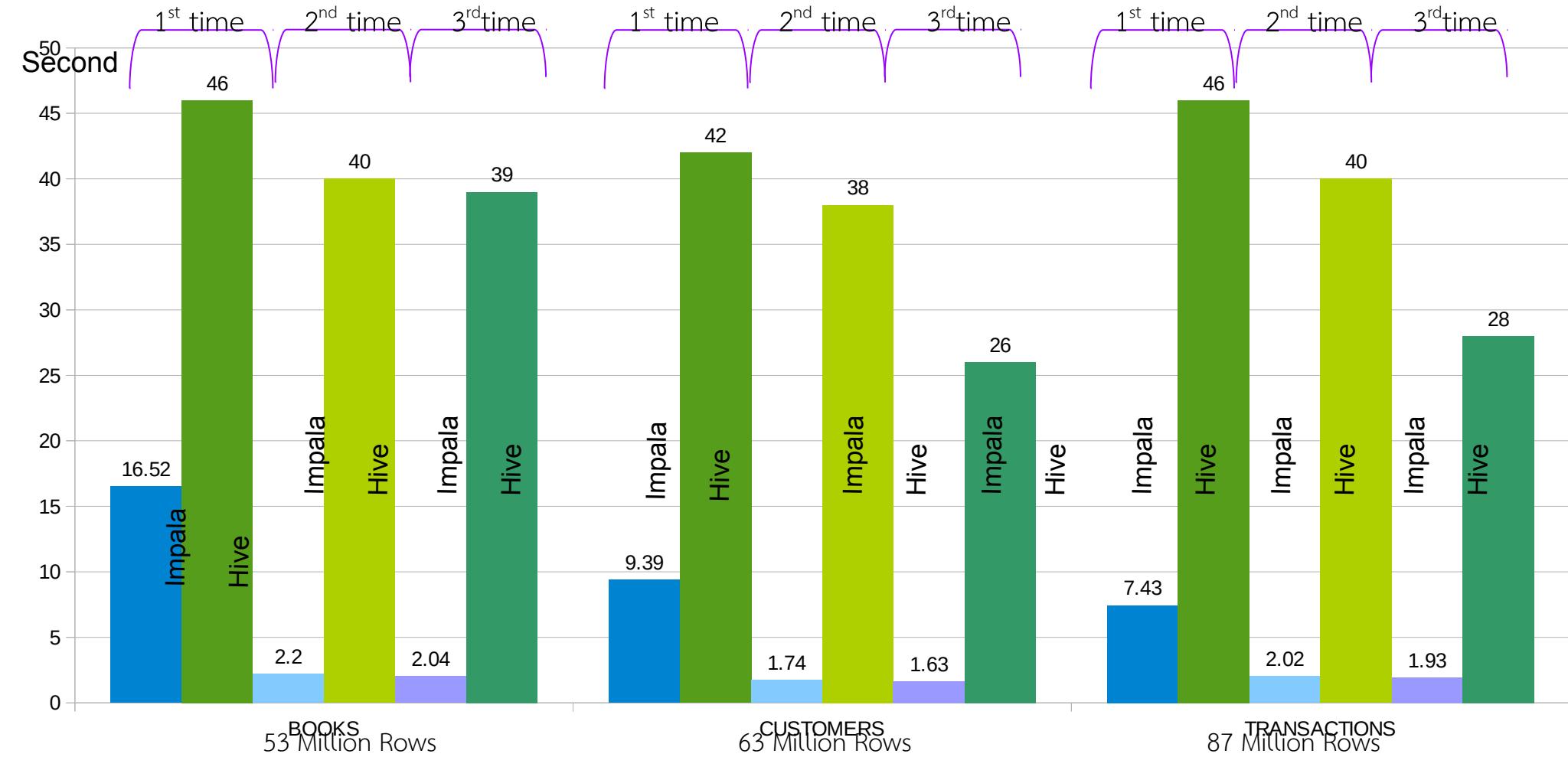
Q1: Scan Query

ค่าเฉลี่ย impala ของ Amazon ที่ 8 Sec / เราก็ 4.98 Sec

ค่าเฉลี่ยของ Hive ของ Amazon ที่ 110 Sec / เราก็ 38.33 Sec

ค่า Second ยิ่งน้อยยิ่งเร็วและดี

| FILESIZE 4GB | | IMPALA QUERY TIME / Second | | | HIVE QUERY TIME / Second | | | QUERY | RESULT |
|--------------|-----------------|----------------------------|----------------------|----------------------|--------------------------|----------------------|----------------------|---|-----------|
| TABLE NAME | TOTAL | 1 st time | 2 nd time | 3 rd time | 1 st time | 2 nd time | 3 rd time | | |
| BOOKS | 53 MILLION ROWS | 16.52 | 2.2 | 2.04 | 46 | 40 | 39 | select COUNT(*) FROM books WHERE category = 'TRAVEL'; | 1,241,166 |
| CUSTOMERS | 63 MILLION ROWS | 9.39 | 1.74 | 1.63 | 42 | 38 | 26 | select COUNT(*) FROM customers WHERE name = 'Harrison SMITH'; | 1,302 |
| TRANSACTIONS | 87 MILLION ROWS | 7.43 | 2.02 | 1.93 | 46 | 40 | 28 | select COUNT(*) FROM transactions WHERE quantity = 15; | 1,778,607 |



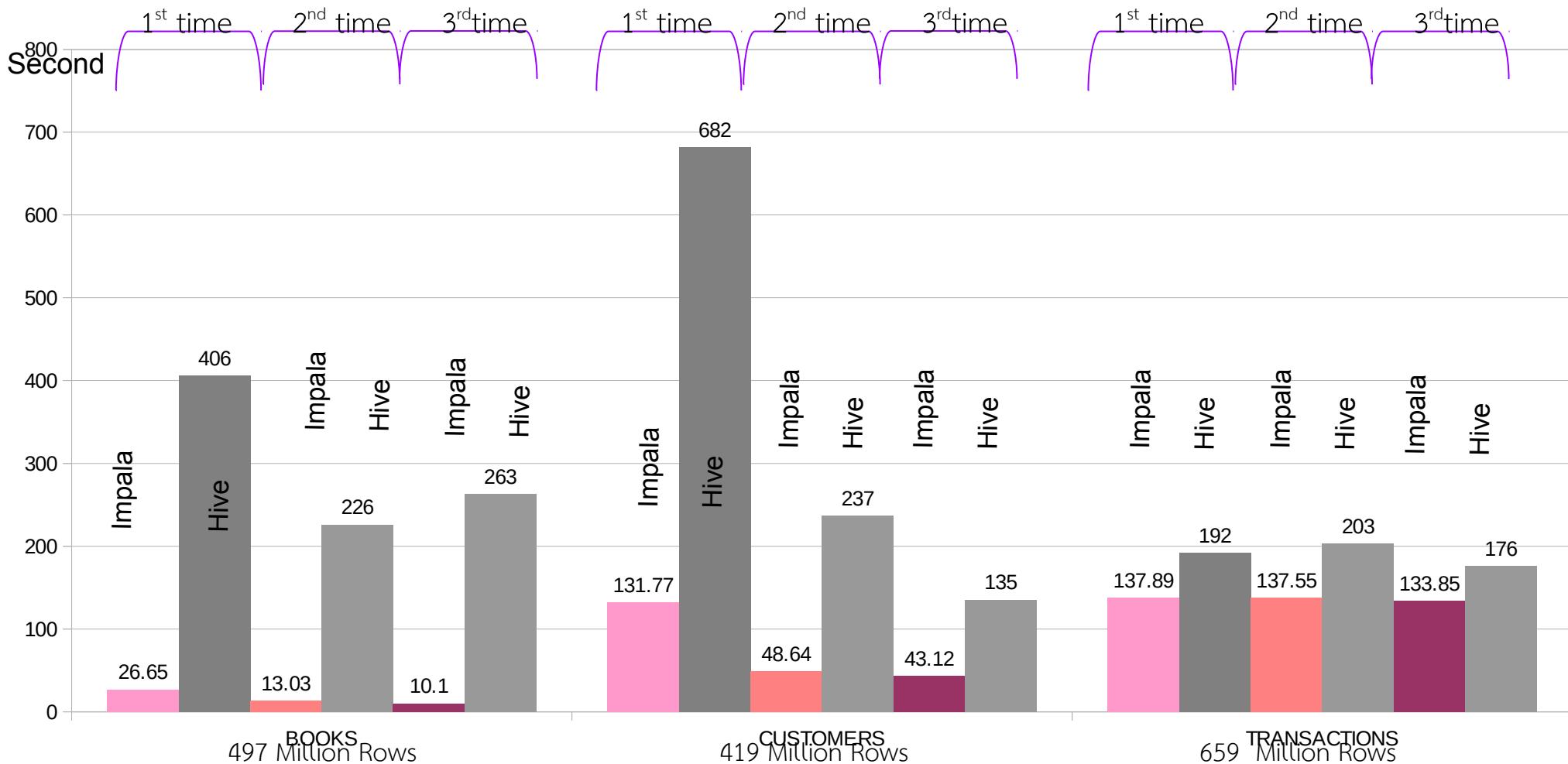
ค่าเฉลี่ย impala ของ Amazon ที่ 82 Sec / เรากำ 75.84 Sec

ค่าเฉลี่ยของ Hive ของ Amazon ที่ 294 Sec / เรากำ 280 Sec

ค่า Second ยิ่งน้อยยิ่งเร็วและดี

Q1: Scan Query

| FILESIZE 32GB | | IMPALA QUERY TIME / Second | | | HIVE QUERY TIME / Second | | | QUERY | RESULT |
|---------------|------------------|----------------------------|----------------------|----------------------|--------------------------|----------------------|----------------------|---|------------|
| TABLE NAME | TOTAL | 1 st time | 2 nd time | 3 rd time | 1 st time | 2 nd time | 3 rd time | | |
| BOOKS | 497 MILLION ROWS | 26.65 | 13.03 | 10.1 | 406 | 226 | 263 | select COUNT(*) FROM books WHERE category = 'TRAVEL'; | 9,803,669 |
| CUSTOMERS | 419 MILLION ROWS | 131.77 | 48.64 | 43.12 | 682 | 237 | 135 | select COUNT(*) FROM customers WHERE name = 'Harrison SMITH'; | 10,636 |
| TRANSACTIONS | 659 MILLION ROWS | 137.89 | 137.55 | 133.85 | 192 | 203 | 176 | select COUNT(*) FROM transactions WHERE quantity = 15; | 13,450,247 |



Q2:Aggregation Query

```
SELECT Field count(*) cnt  
FROM Table  
GROUP BY Field  
ORDER BY cnt DESC LIMIT  
10;
```

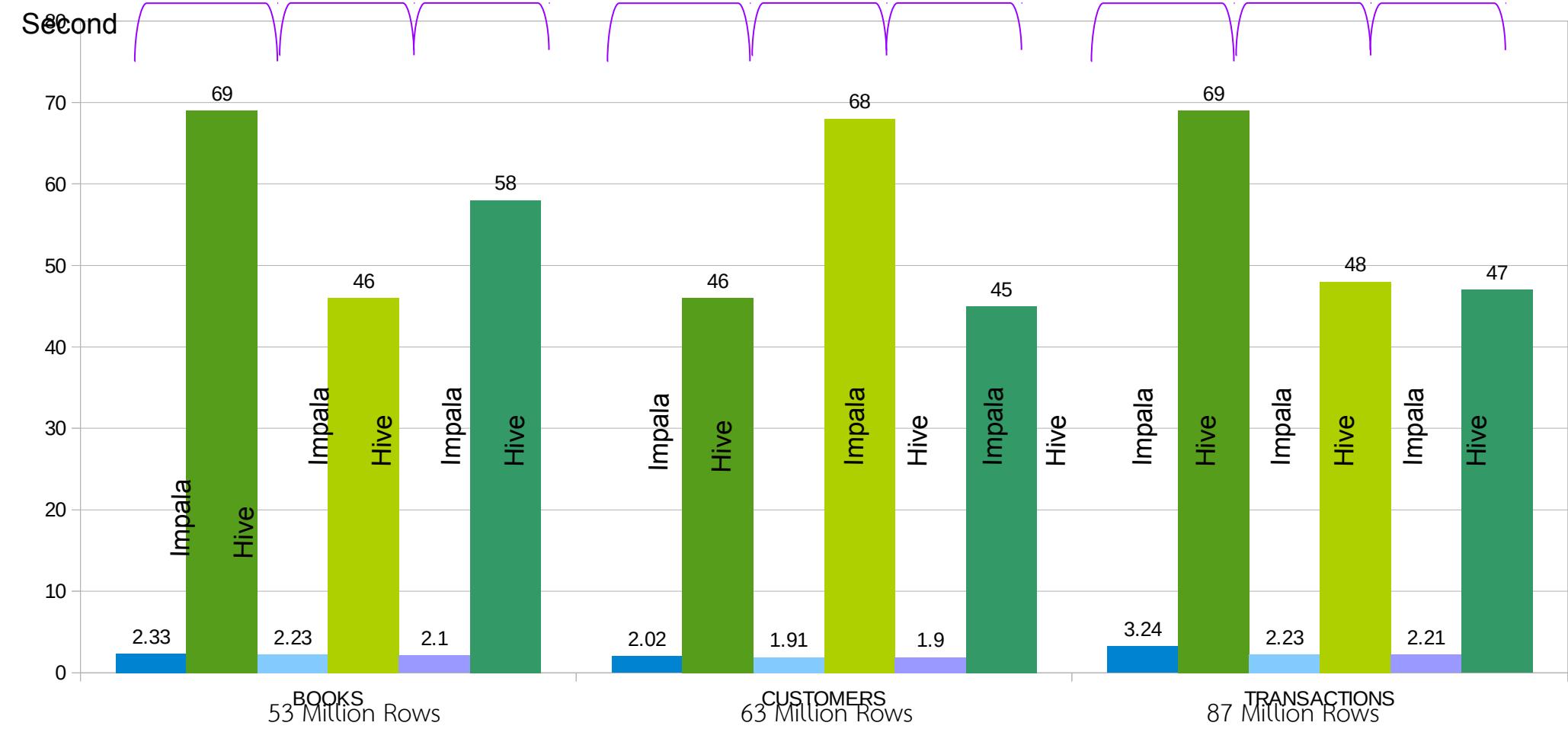
The aggregation query scans a single table, groups the rows, and calculates the size of each group

Q2: Aggregation Query

ค่าเฉลี่ย impala ของ Amazon ที่ 18 Sec / เรากำ 2.24 Sec

ค่าเฉลี่ยของ Hive ของ Amazon ที่ 130 Sec / เรากำ 55.11 Sec คำ Second ยังน้อยกว่าเราและดี

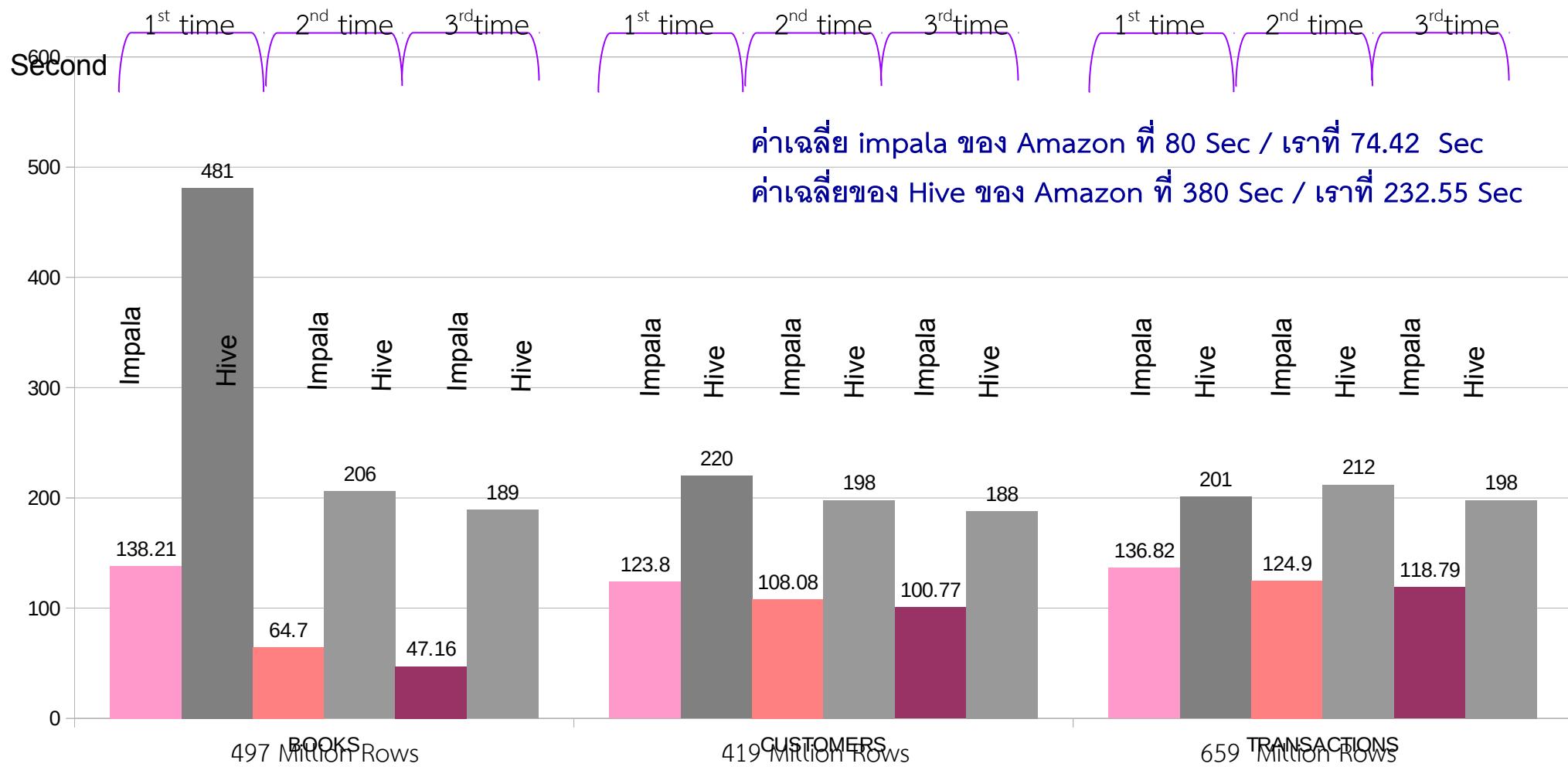
| FILESIZE 4GB | | IMPALA QUERY TIME / Second | | | HIVE QUERY TIME / Second | | | QUERY |
|--------------|-----------------|----------------------------|----------------------|----------------------|--------------------------|----------------------|----------------------|---|
| TABLE NAME | TOTAL | 1 st time | 2 nd time | 3 rd time | 1 st time | 2 nd time | 3 rd time | |
| BOOKS | 53 MILLION ROWS | 2.33 | 2.23 | 2.1 | 69 | 46 | 58 | Select category, count(*) cnt FROM books GROUP BY category ORDER BY cnt DESC LIMIT 10; |
| CUSTOMERS | 63 MILLION ROWS | 2.02 | 1.91 | 1.9 | 46 | 68 | 45 | Select state, count(*) cnt FROM customers GROUP BY state ORDER BY cnt DESC LIMIT 10; |
| TRANSACTIONS | 87 MILLION ROWS | 3.24 | 2.23 | 2.21 | 69 | 48 | 47 | Select quantity, count(*) cnt FROM transactions GROUP BY quantity ORDER BY cnt DESC LIMIT 10; |



O2: Aggregation Query

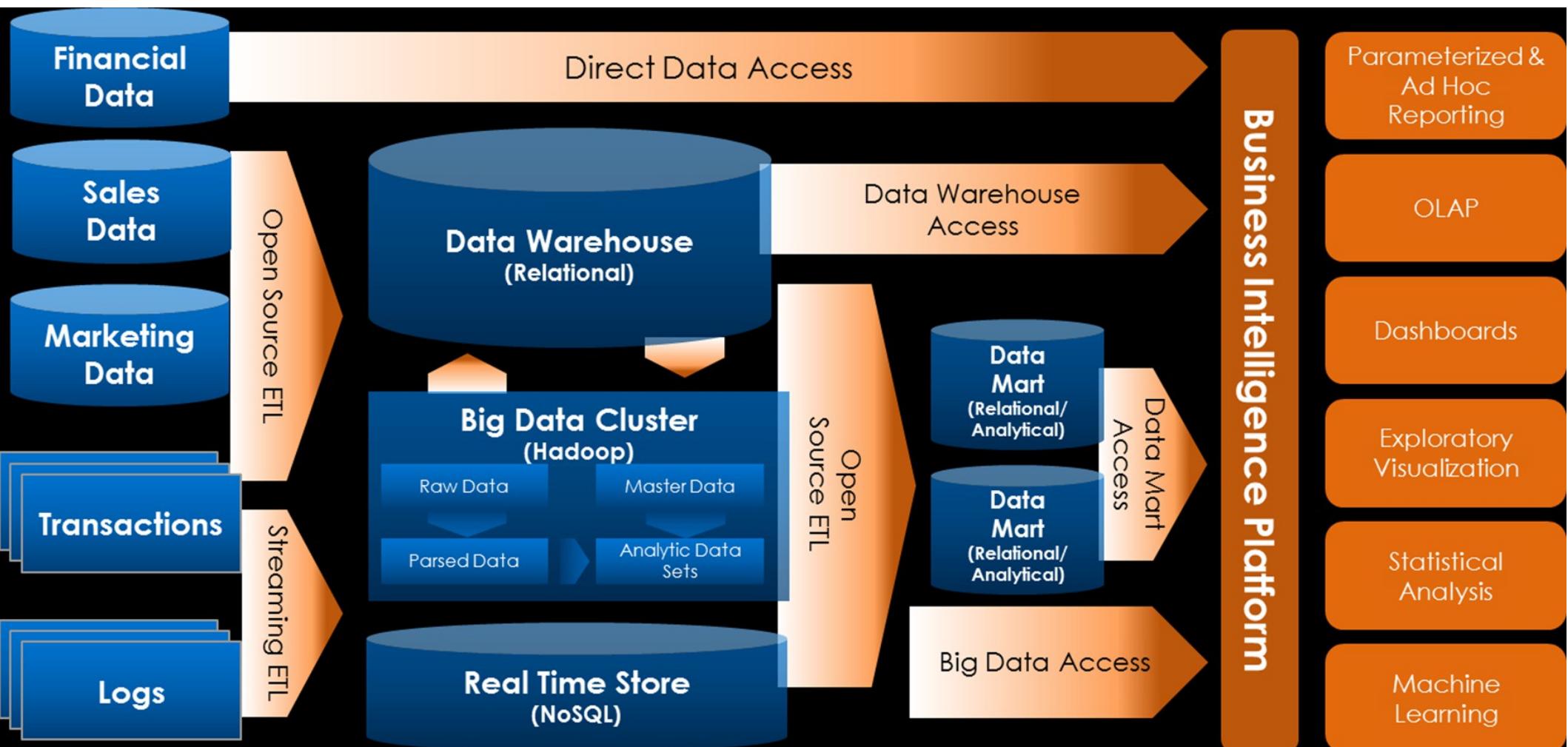
ค่า Second ยิงน้อยยิงเร็วและดี

| FILESIZE 32GB | | IMPALA QUERY TIME / Second | | | HIVE QUERY TIME / Second | | | QUERY |
|---------------|------------------|----------------------------|----------------------|----------------------|--------------------------|----------------------|----------------------|---|
| TABLE NAME | TOTAL | 1 st time | 2 nd time | 3 rd time | 1 st time | 2 nd time | 3 rd time | |
| BOOKS | 497 MILLION ROWS | 138.21 | 64.7 | 47.16 | 481 | 206 | 189 | Select category, count(*) cnt FROM books GROUP BY category ORDER BY cnt DESC LIMIT 10; |
| CUSTOMERS | 419 MILLION ROWS | 123.8 | 108.08 | 100.77 | 220 | 198 | 188 | Select state, count(*) cnt FROM customers GROUP BY state ORDER BY cnt DESC LIMIT 10; |
| TRANSACTIONS | 659 MILLION ROWS | 136.82 | 124.9 | 118.79 | 201 | 212 | 198 | Select quantity, count(*) cnt FROM transactions GROUP BY quantity ORDER BY cnt DESC LIMIT 10; |



BIGDATA

แล้วไปไหน



TRIED AND TRUE METHODS FOR APPLYING BIG DATA



KEEP A CLOSE EYE ON SOCIAL MEDIA
by tracking your hits, user base, and amount of shares



LEARN HOW TO READ A P&L
Use to ensure profitable growth and expansion



ANALYZE THE EFFECTIVENESS OF YOUR WEBSITE AND MARKETING
Users visited since latest marketing ploy, users lost, and increased revenue



MONITOR YOUR CONNECTIONS ON SOCIAL NETWORKS
Find your competitors, and see how you compare



USE GOOGLE TRENDS
To explore relevant keywords and apply to your business to increase revenue and traffic



USE QUANTCAST
To 'quantify' your website -- daily hits, devices used (mobile, laptop, etc.), gender, location, and age -- then customize based on data trends



REVIEW WEBSITE ANALYTICS TO FIND THE AVERAGE VISITOR TIME SPENT
Too short and it's likely not engaging; too long and your site may need to be more condensed/intuitive

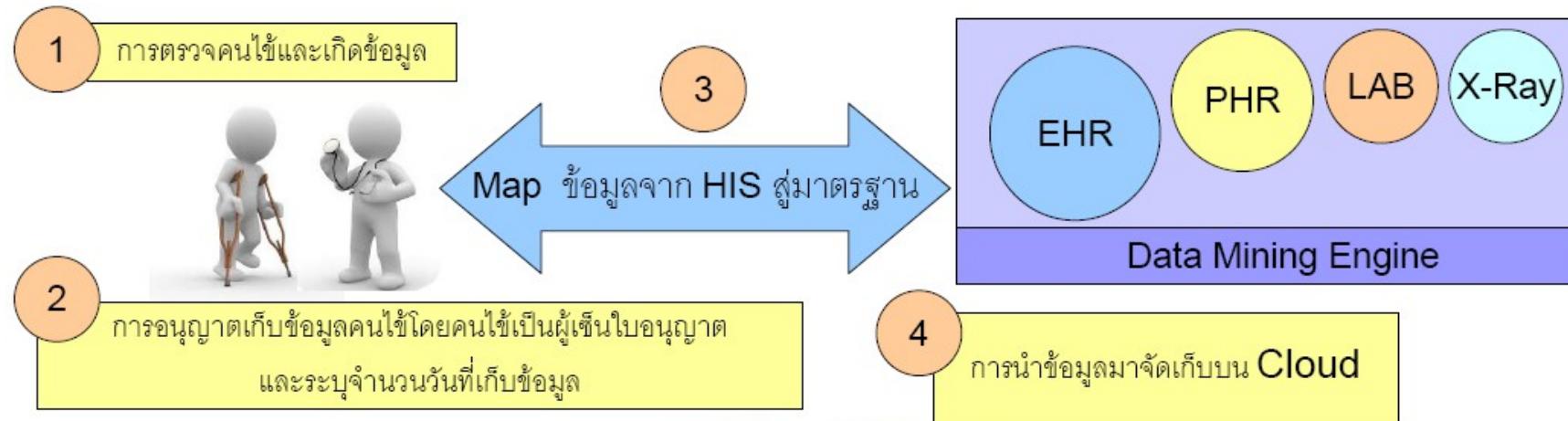


RETARGET
By shifting your brand or marketing efforts



GET TO KNOW YOUR AUDIENCE/CUSTOMER
By analyzing their spending habits

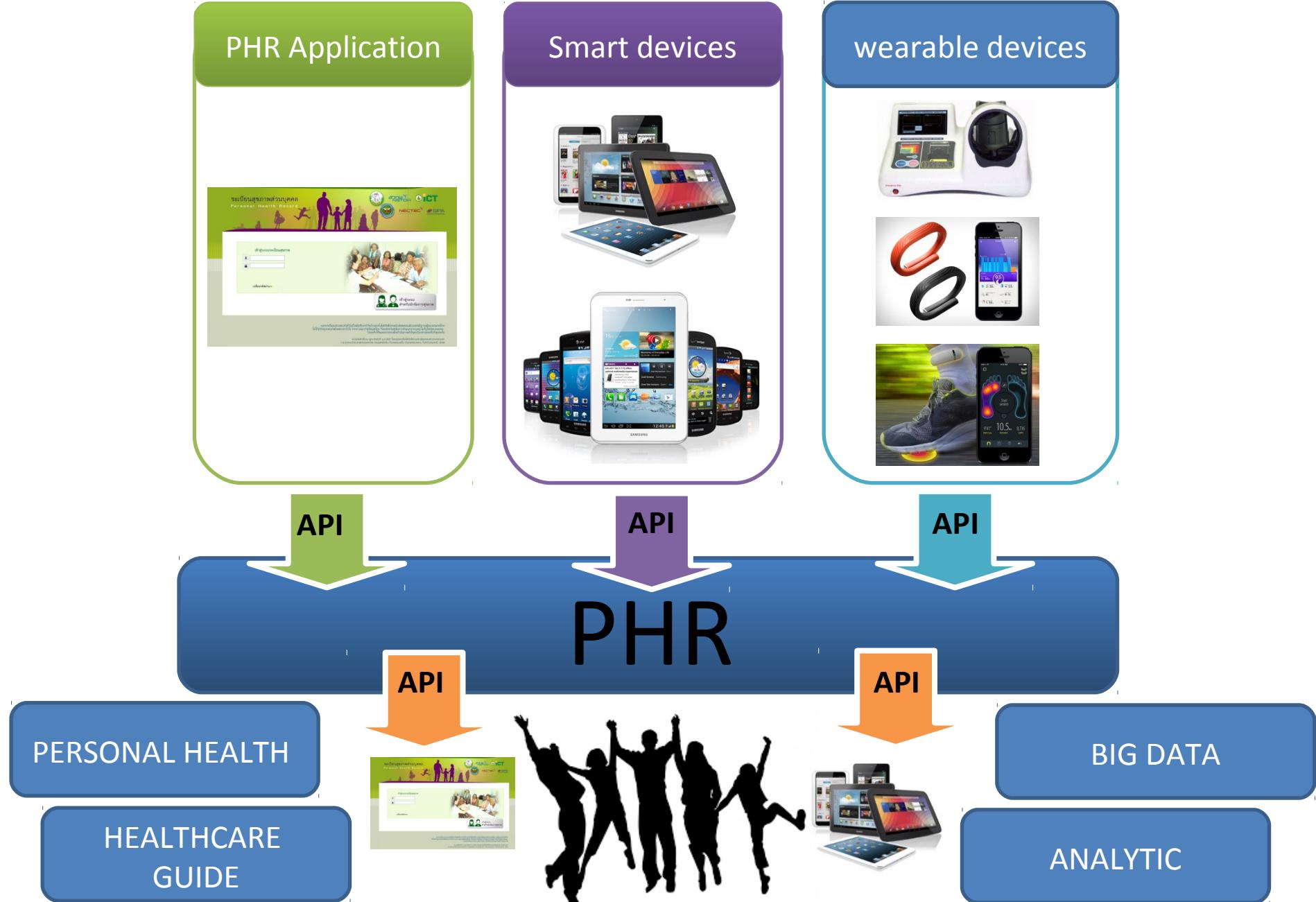
Health Care Concept

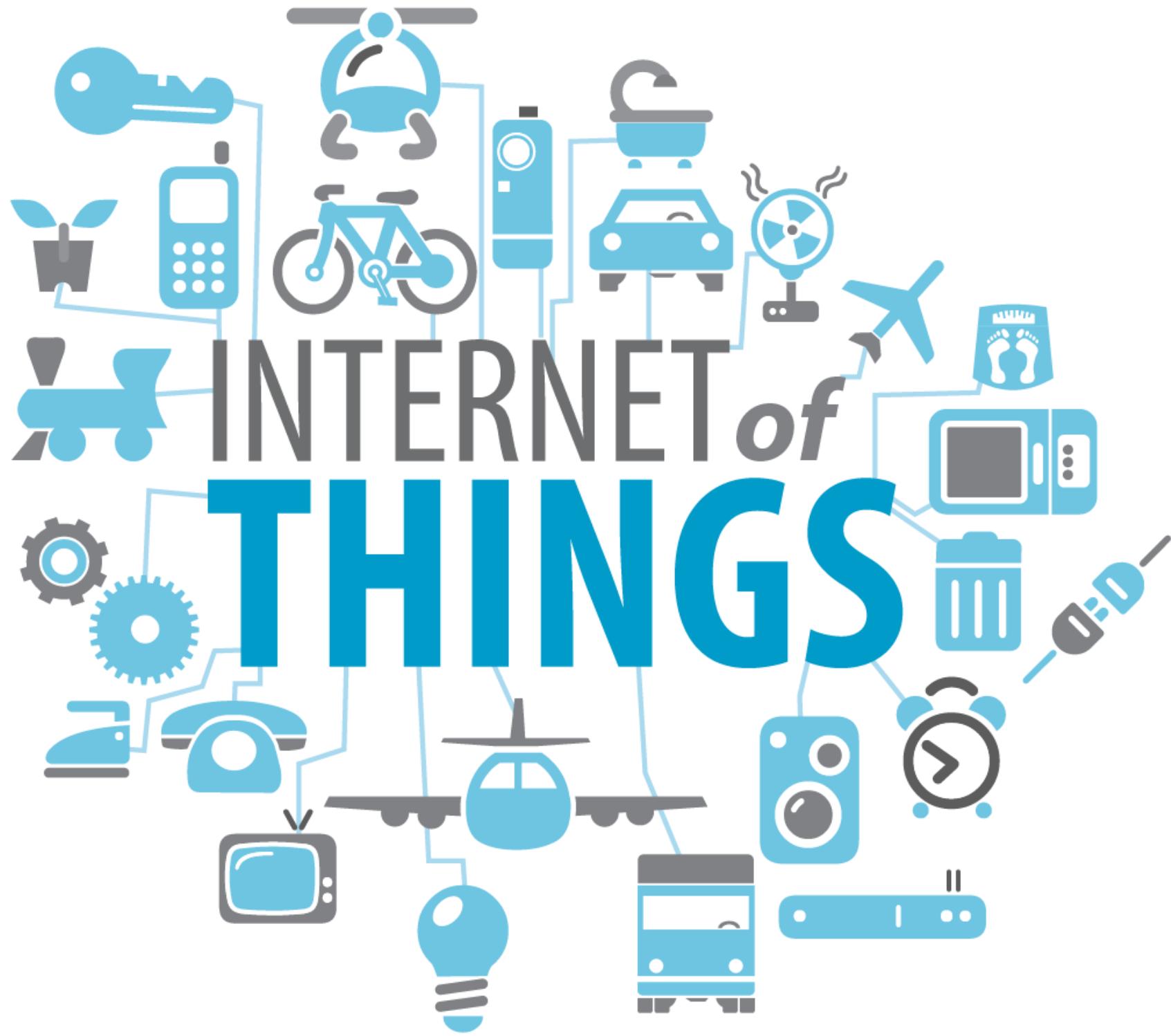


| |
|--------------------------|
| Usability |
| - Dropbox Feature |
| - Offline Access |
| - Device Compatible |
| Speed Performance |
| - Bandwidth |
| - UL/DL Speed |
| - Private Bandwidth |
| Security |
| - Data Encrypt |
| - Key Authentication |
| - https |
| - VPN |



เทคโนโลยีการบันทึกข้อมูลสุขภาพผ่านอุปกรณ์อัจฉริยะที่สามารถเชื่อมโยงผ่าน PHR API





แล้ววันนี้คุณเลือกหรือยัง ว่าจะอยู่จุดไหนของ BIG DATA ?

Q&A

Thankyou