MONASH University

MONASH
INFORMATION
TECHNOLOGY

Database Current Trends
Exam Preparation

GROUP OF EIGHT AUSTRALIA

Week 12 Unit Test

- TIME and LOCATION: Laboratory class, week 12
  - You **MUST** attend test in your allocated lab.
  - You **MUST NOT** be late. If you are late, you will not be allowed in the lab and will receive zero mark.
- DURATION: 90 minutes (including submission time)
- TOPIC: Writing SQL to retrieve data. (week 7, 9 and10)
- Test to be conducted on BUGS database
  - **DO NOT** run the schema file in your account. If you have run, delete it from the account.
  - Look at the comments column if unsure what a column means
- Don’t worry about the efficiency of SQL query as long as it is correct
- During your test
  - Submit to Moodle as you go through the test (so that you don’t lose any work).
  - Before uploading the file to Moodle, save & close the file in SQL developer and upload it. Uploading without saving and closing it in SQL developer may upload an empty file.
  - 15-30 minutes before the end of the test, download the file from Moodle to make sure it is the right one

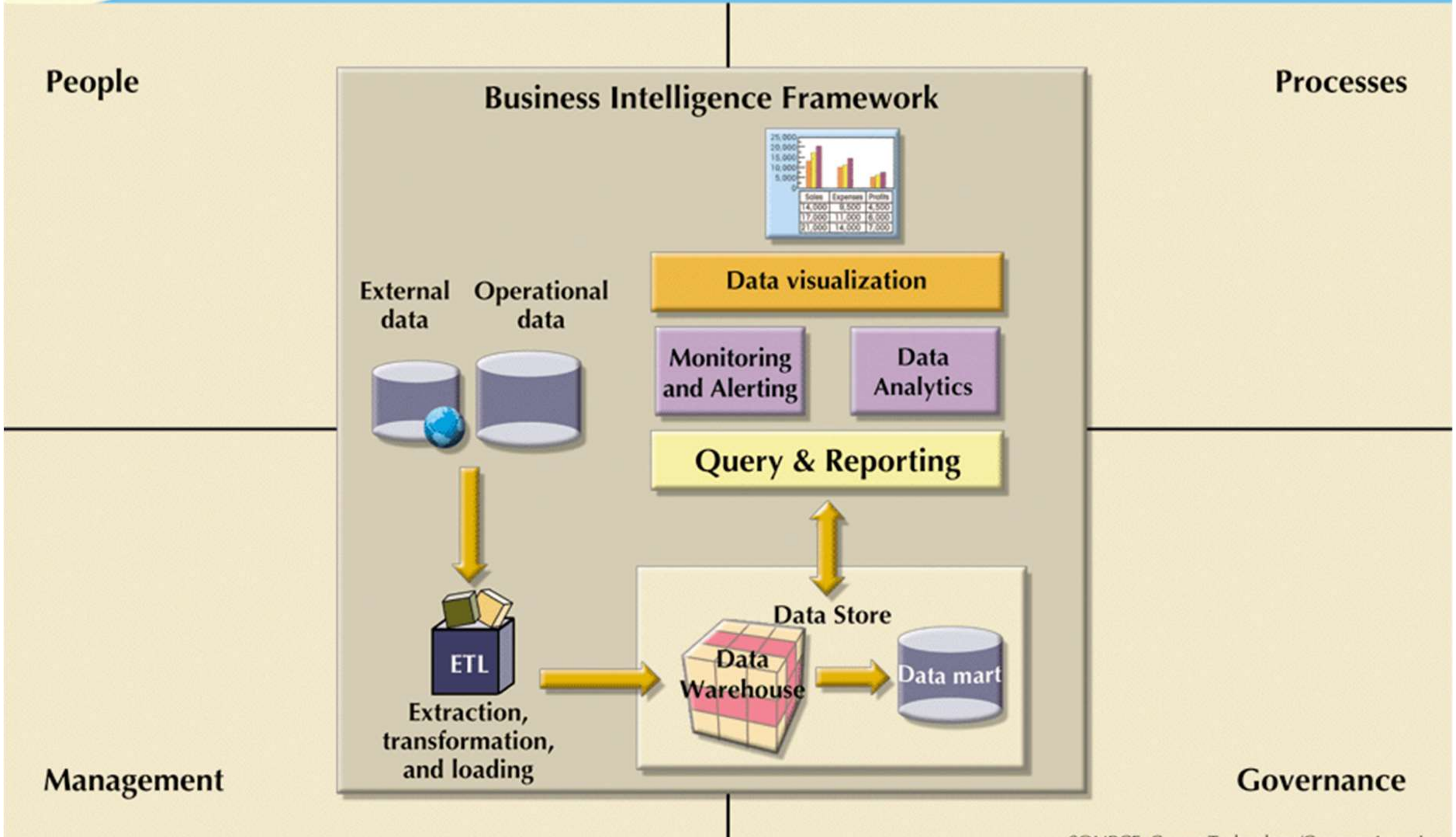
Usage of database

- Example of a supermarket
- Decision making
  - Operational level
    - When do we need to re-stock X-item?
  - Strategic and tactical level
    - Is there any branch that performs worse than the state average?
    - What is the total sales made by each state each year and across a number of years?
    - What a particular customer is interested in or would be interested in near-future?

Operational Data vs. Decision Support Data

- Operational data
  - Mostly stored in relational database
  - Optimized to support transactions representing daily operations
  - Example:
    - How many students enrolled in FIT2094?
- Decision support data differs from operational data in three main areas:
  - Time span
  - Granularity
  - Dimensionality
  - Example:
    - What is the total number of students in the foundation units in each year (subtotal of the two semesters numbers) and the total across years, across a single unit.

FIGURE 13.1 Business intelligence framework



SOURCE: Course Technology/Cengage Learning

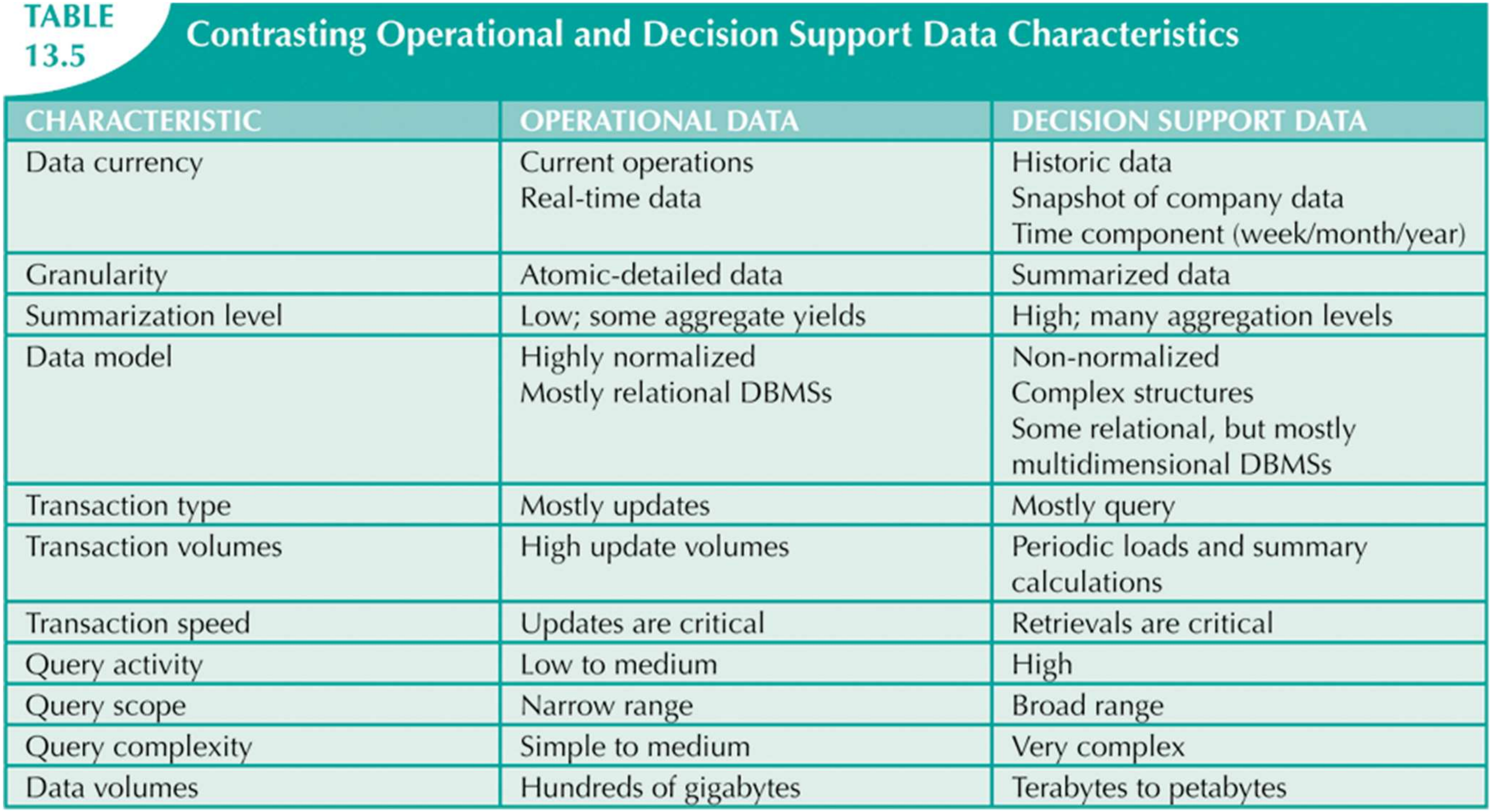


TABLE 13.5 Contrasting Operational and Decision Support Data Characteristics

CHARACTERISTIC	OPERATIONAL DATA	DECISION SUPPORT DATA
Data currency	Current operations Real-time data	Historic data Snapshot of company data Time component (week/month/year)
Granularity	Atomic-detailed data	Summarized data
Summarization level	Low; some aggregate yields	High; many aggregation levels
Data model	Highly normalized Mostly relational DBMSs	Non-normalized Complex structures Some relational, but mostly multidimensional DBMSs
Transaction type	Mostly updates	Mostly query
Transaction volumes	High update volumes	Periodic loads and summary calculations
Transaction speed	Updates are critical	Retrievals are critical
Query activity	Low to medium	High
Query scope	Narrow range	Broad range
Query complexity	Simple to medium	Very complex
Data volumes	Hundreds of gigabytes	Terabytes to petabytes

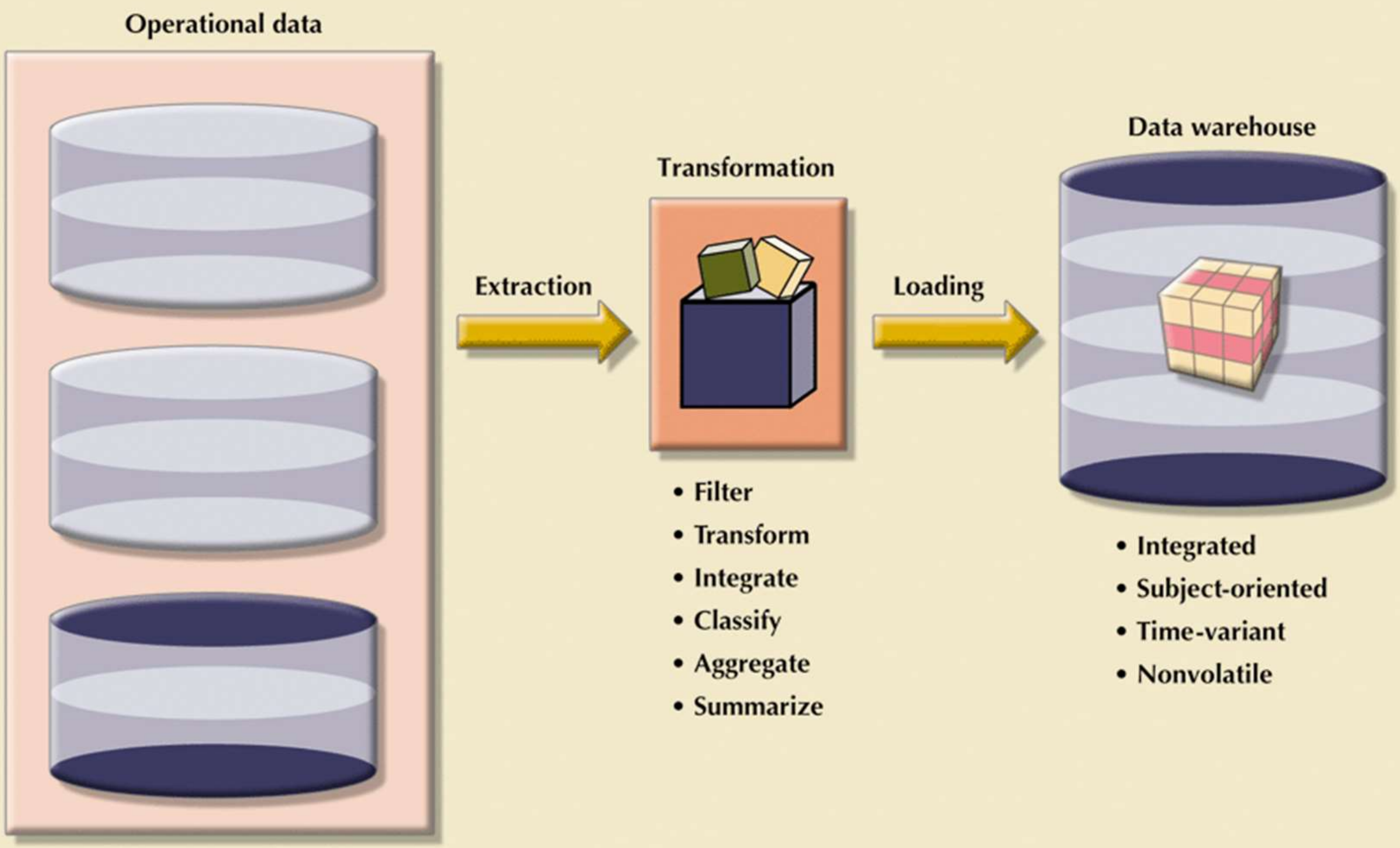
Decision Support Database Requirements

- Specialized DBMS tailored to provide fast answers to complex queries
- Three main requirements
  - Database schema
  - Data extraction and loading
  - Database size
- Database schema
  - Complex data representations
  - Aggregated and summarized data
  - Queries extract multidimensional time slices
- Data extraction and filtering
  - Supports different data sources
    - Flat files
    - Hierarchical, network, and relational databases
    - Multiple vendors
  - Checking for inconsistent data

The Data Warehouse (FIT3003, FIT5195, FIT5137)

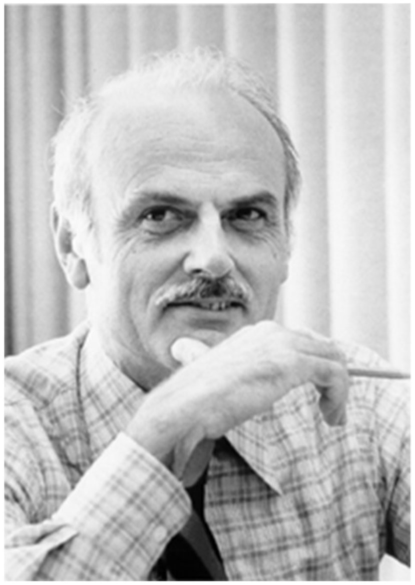
- Database size
  - In 2013, eBay had around 50 Petabytes of data in its data warehouses (50,000 Terabytes)
  - DBMS must support very large databases (VLDBs)
- Integrated, subject-oriented, time-variant, and nonvolatile collection of data
  - Provides support for decision making
- Usually a read-only database optimized for data analysis and query processing
- Requires time, money, and considerable managerial effort to create

FIGURE 13.4 The ETL process



SOURCE: Course Technology/Cengage Learning

Database Hall of Fame



E.F “Ted” Codd

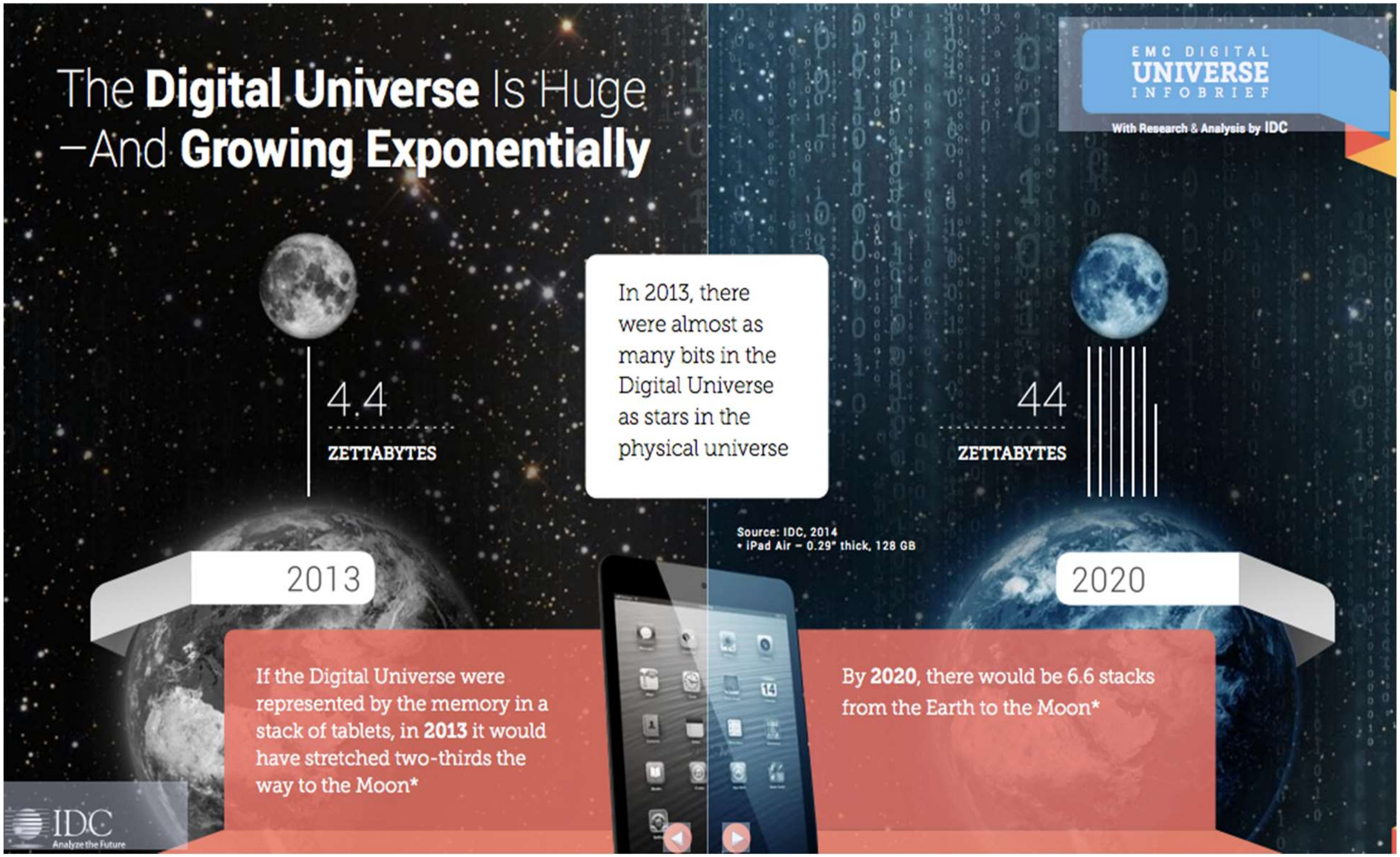
Larry Ellison

Peter Chen

Michael Stonebraker

Internet of Things (IoT)

What is happening with data?



The Digital Universe Is Huge —And Growing Exponentially

In 2013, there were almost as many bits in the Digital Universe as stars in the physical universe

If the Digital Universe were represented by the memory in a stack of tablets, in **2013** it would have stretched two-thirds the way to the Moon*

By **2020**, there would be 6.6 stacks from the Earth to the Moon*

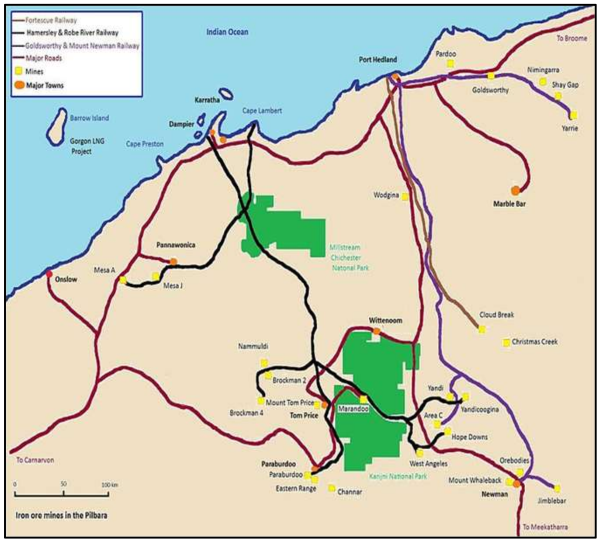
Source: IDC, 2014
• iPad Air – 0.29" thick, 128 GB

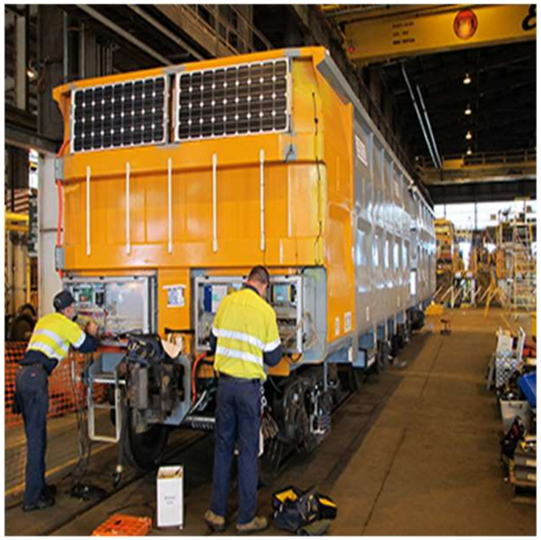
EMC DIGITAL UNIVERSE INFOBRIEF
With Research & Analysis by IDC
4.4
ZETTABYTES
2013
44
ZETTABYTES
2020
IDC
Analyze the Future

1 ZB = 10^{21} bytes = 1 billion terabytes = 1 trillion gigabytes

http://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf

Railway In Mining



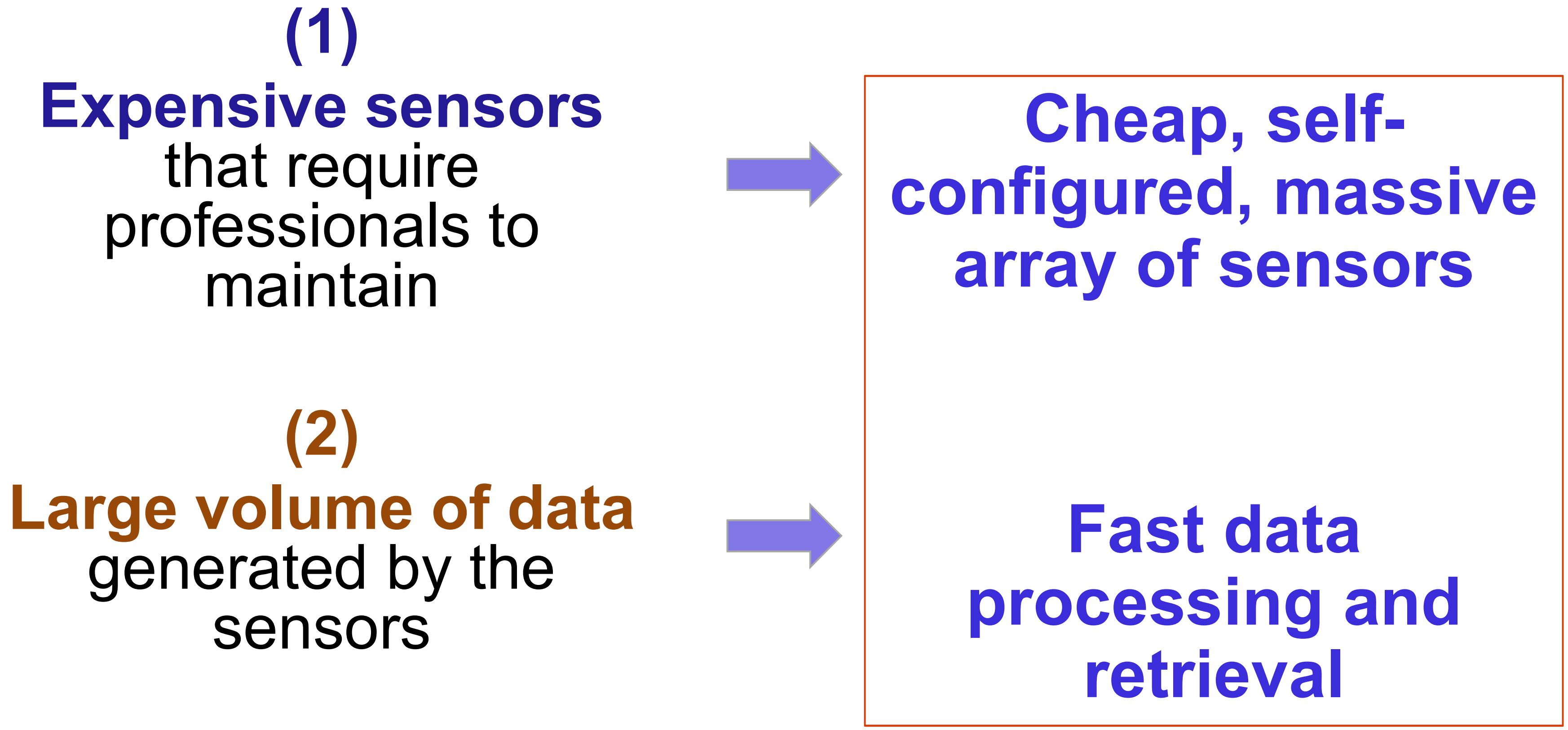


- Pilbara region, WA
- Trains perform round trips from the mining site to the port
- Loaded minerals and ores

- Length: > 2KM
- Load: > 10 Ton/car
- Speed: 5-10 Km/hr

- Instrumented Ore Car (IOC)
- Expensive Sensors
- Trained Professionals to maintain the sensors

Challenges



(1)

Expensive sensors that require professionals to maintain

Cheap, self-configured, massive array of sensors

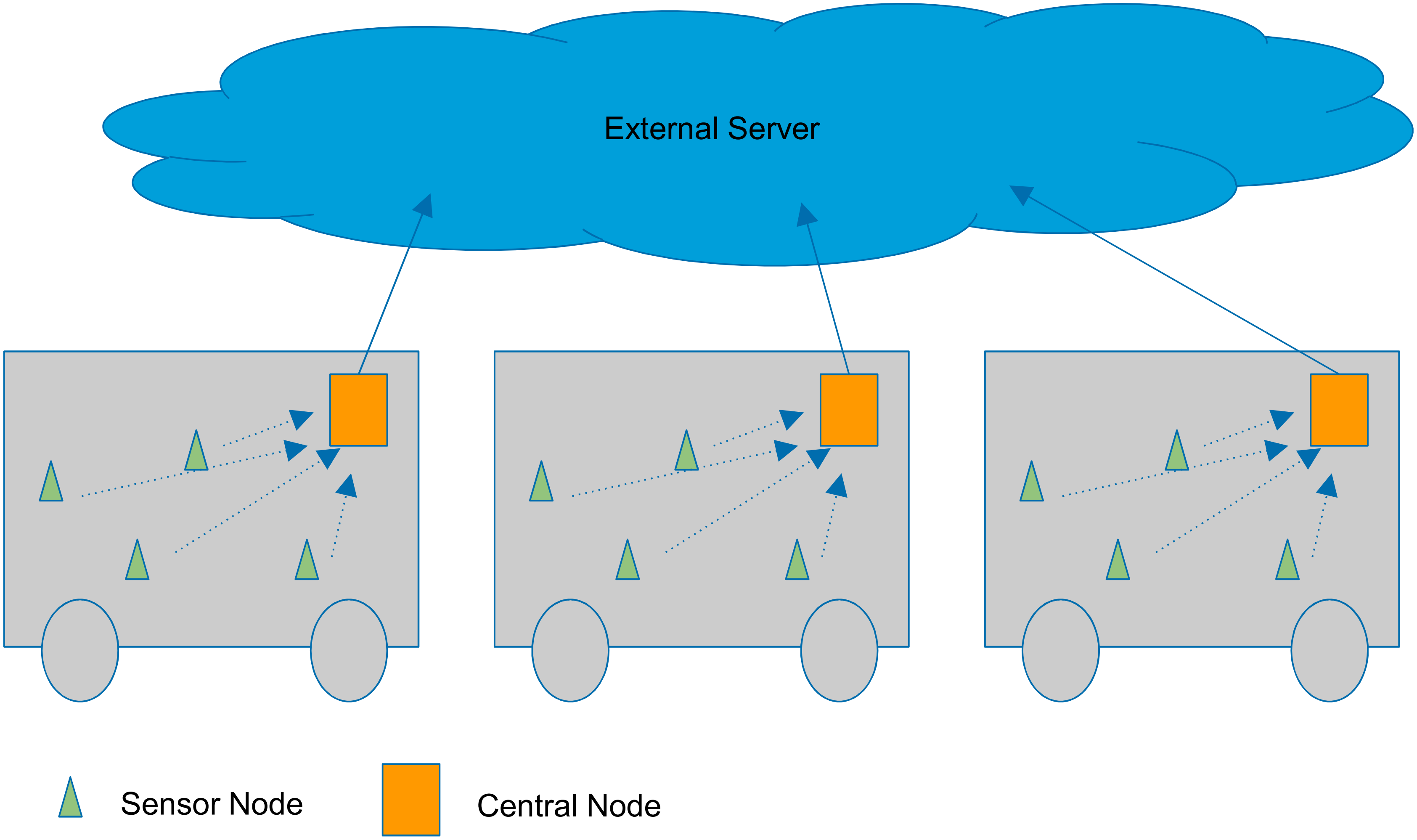
(2)

Large volume of data generated by the sensors

Fast data processing and retrieval

Needs expertise from Eng and IT

Network Structure

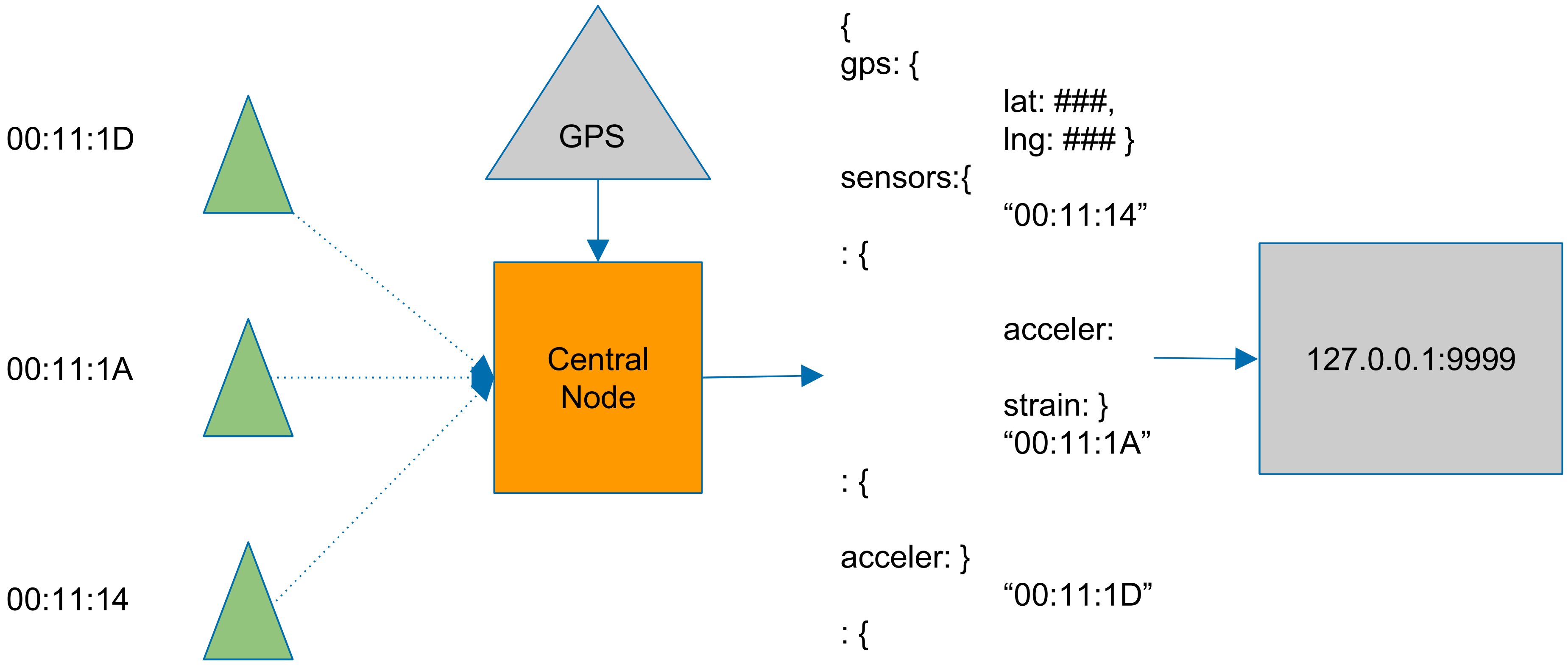


External Server

Sensor Node

Central Node

Central Node Process



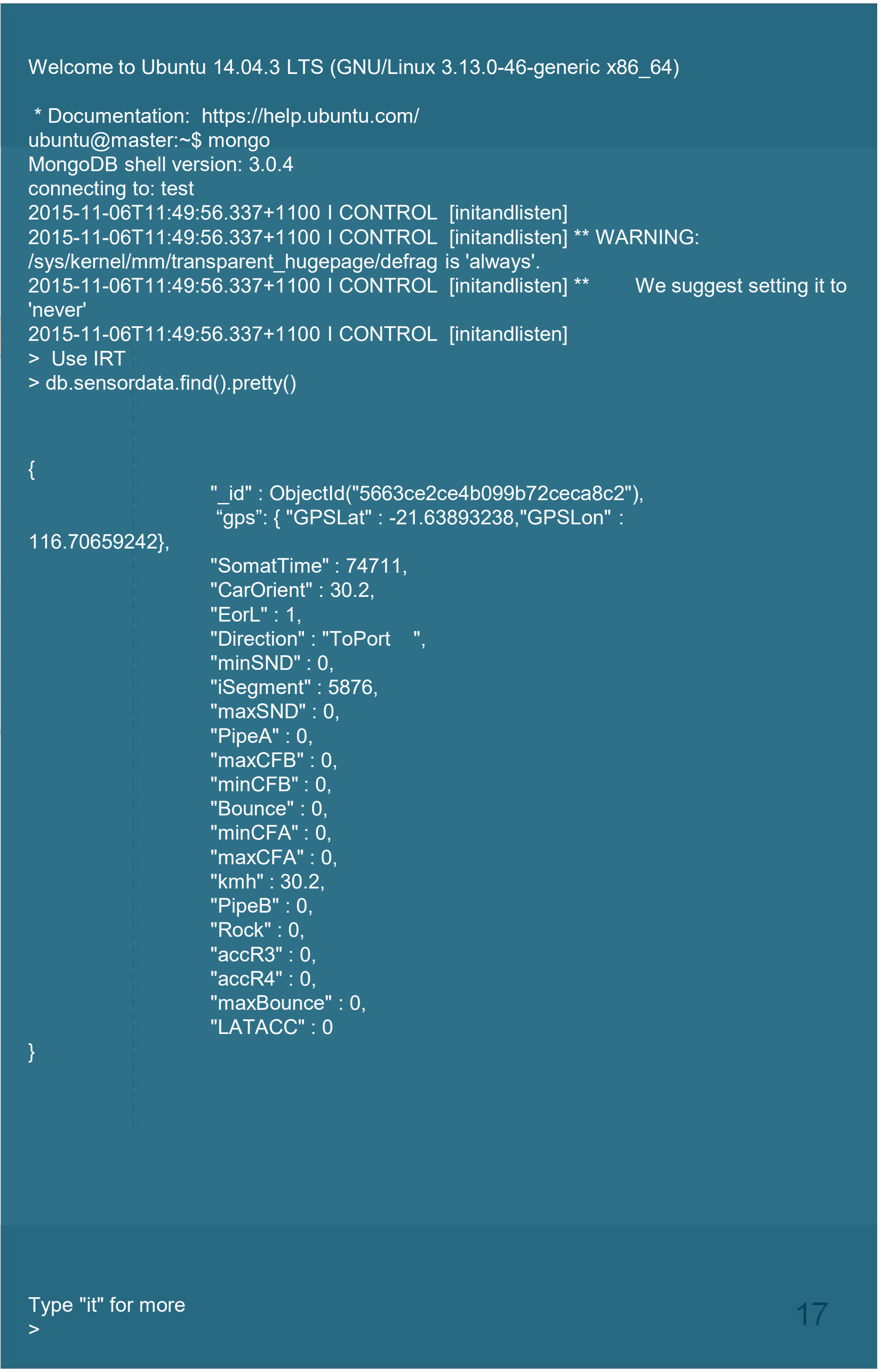
```
{
gps: {
                lat: ###,
                lng: ### }
sensors:{
                “00:11:14”
: {

                acceler:

                strain: }
                “00:11:1A”
: {

acceler: }
                “00:11:1D”
: {

                accler: }}
}
```

How Big is the Data?

Quantity	Data Returned
Timestamp	12-Jun-2015; 09:35:15
Geo-location	N35˚43.57518,W078˚49.78314
Direction	ToMine
Acceleration	0.285g
Pressure	65psi
Ambient temperature	73 degrees F
Surface temperature	78 degrees F
Humidity	35%

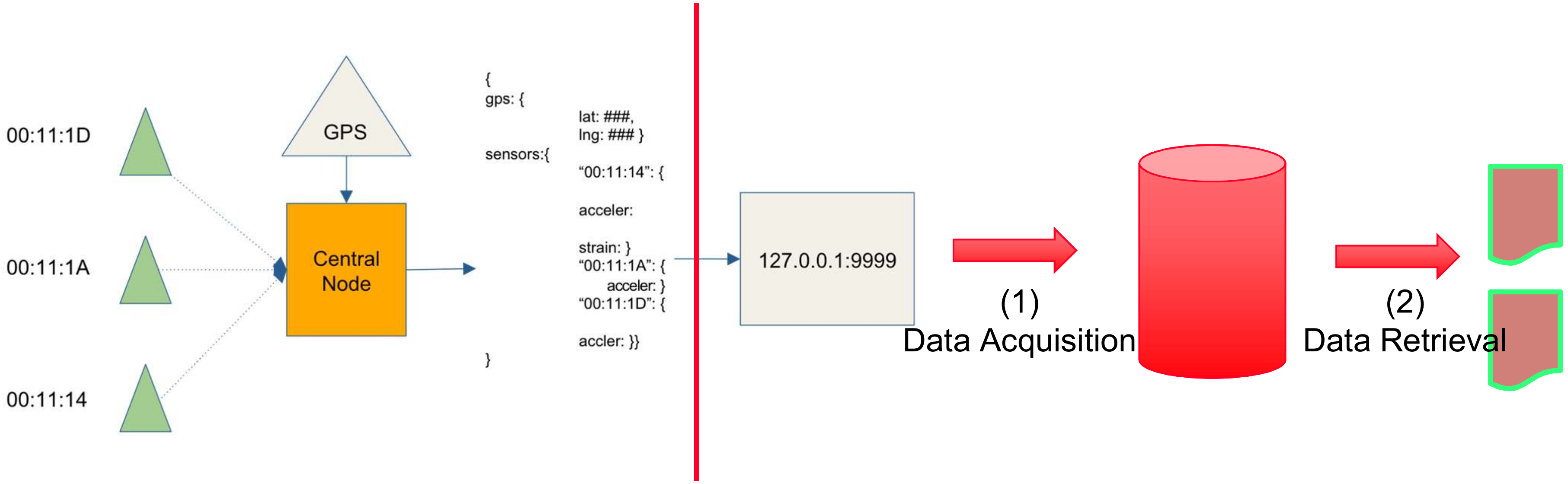
- 16 Sensors
- 200 Ore Cars
- 25 Records Per Second

16 * 200 * 25 = 80,000 records/sec



```
Welcome to Ubuntu 14.04.3 LTS (GNU/Linux 3.13.0-46-generic x86_64)

 * Documentation:  https://help.ubuntu.com/
ubuntu@master:~$ mongo
MongoDB shell version: 3.0.4
connecting to: test
2015-11-06T11:49:56.337+1100 I CONTROL  [initandlisten]
2015-11-06T11:49:56.337+1100 I CONTROL  [initandlisten] ** WARNING:
/sys/kernel/mm/transparent_hugepage/defrag is 'always'.
2015-11-06T11:49:56.337+1100 I CONTROL  [initandlisten] **        We suggest setting it to
'never'
2015-11-06T11:49:56.337+1100 I CONTROL  [initandlisten]
>  Use IRT
> db.sensordata.find().pretty()


{
                "_id" : ObjectId("5663ce2ce4b099b72ceca8c2"),
                 "gps": { "GPSLat" : -21.63893238,"GPSLon" :
116.70659242},
                "SomatTime" : 74711,
                "CarOrient" : 30.2,
                "EorL" : 1,
                "Direction" : "ToPort    ",
                "minSND" : 0,
                "iSegment" : 5876,
                "maxSND" : 0,
                "PipeA" : 0,
                "maxCFB" : 0,
                "minCFB" : 0,
                "Bounce" : 0,
                "minCFA" : 0,
                "maxCFA" : 0,
                "kmh" : 30.2,
                "PipeB" : 0,
                "Rock" : 0,
                "accR3" : 0,
                "accR4" : 0,
                "maxBounce" : 0,
                "LATACC" : 0
}


Type "it" for more
>
```

Big Data Processing



```
{
gps: {
                lat: ###,
                lng: ### }
sensors:{
                “00:11:14”: {

                acceler:

                strain: }
                “00:11:1A”: {
                      acceler: }
                “00:11:1D”: {

                accler: }}
}
```

Two main problems:

1. How to receive data … massive amount of data
2. How to retrieve data … very fast

Scaling

- How do we scale current relational systems?

- SQL designed for database as a single physical entity
  - Purchase bigger "boxes": costly and has real limits
  - Increase the number of processors, yielding parallel computation/database with complex issues to handle
  - Distribute database – challenges to maintain ACID transaction principles and issues of availability/consistency

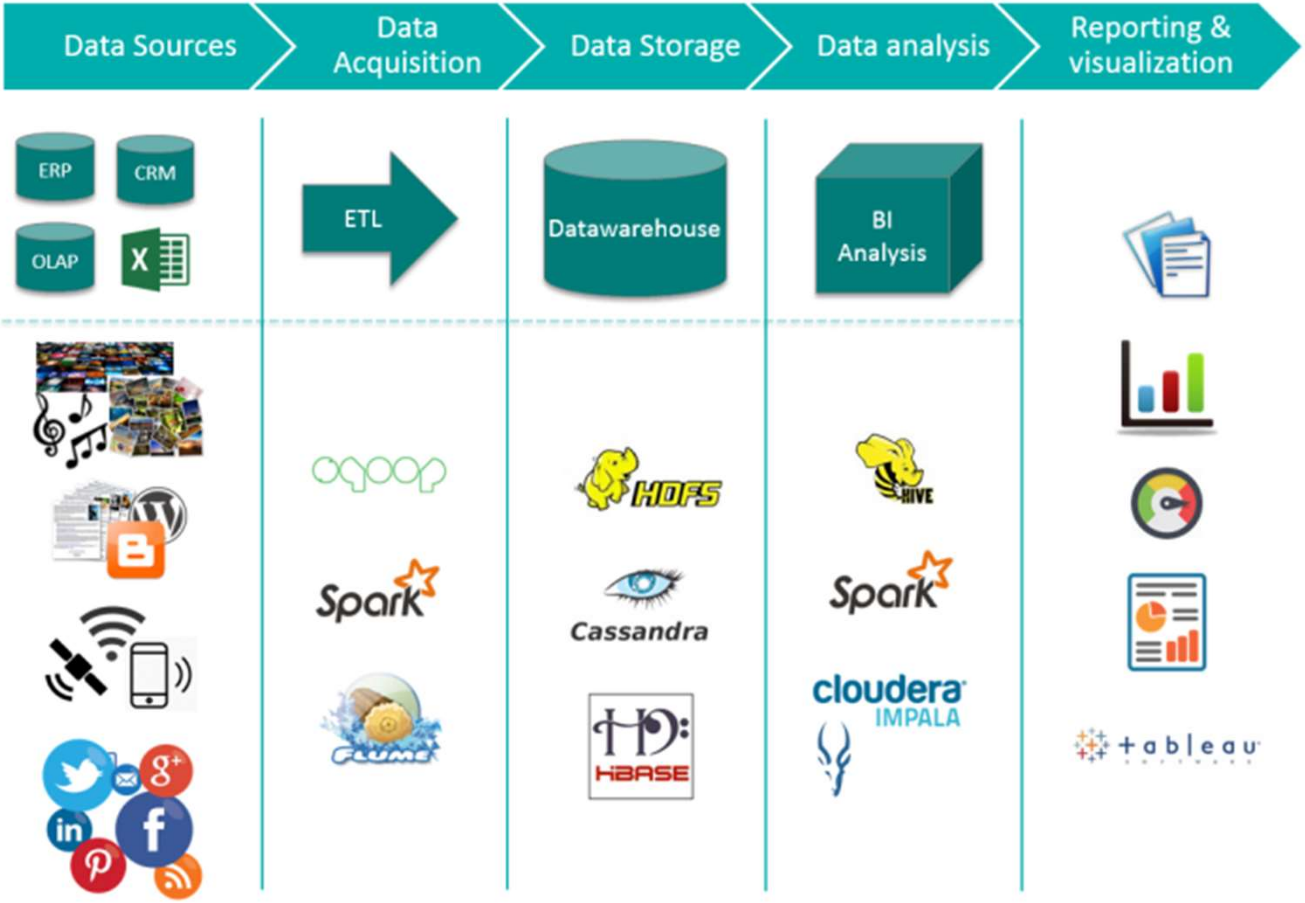
Scaling continued

- Big players, notably Google and Amazon chose a different path
  - Lots and lots of smaller boxes ("commodity" servers)
  - Non relational structure
  - Google: Bigtable
    - http://static.googleusercontent.com/media/research.google.com/en//archive/bigtable-osdi06.pdf
  - Amazon: Dyanmo
    - http://www.read.seas.harvard.edu/~kohler/class/cs239-w08/decandia07dynamo.pdf
- Term "NoSQL" coined by John Oskarsson in 2009 after calling a ..."free meetup about “open source, distributed, non relational databases” or NOSQL for short"…
  - http://blog.oskarsson.nu/post/22996139456/nosql-meetup
- Characteristics
  - Non relational, mostly open source, distributed (cluster friendly), schema-less (no fixed storage schema)
  - See MongoDB "NoSQL Databases Explained"

Fast Data Processing (FIT5202, FIT5148)

- Computer systems
  - –Parallel computer
    - •A single machine with massive number of CPUs.
  - –Cluster of computers
    - •Multiple machines connected via network.
    - •Commodity computer.
- Database structure
  - –Non-relational database (NoSQL)
    - •No update, append only.
    - •Optimised for a ’main’ operation
    - •Examples: MongoDB, Cassandra
  - –Distributed File Systems
    - •HDFS (Hadoop File Systems)
    - •Parquee File Systems
- Parallel data processing
  - –Hadoop
  - –Spark
- In Memory database

Data Processing Ecosystem



http://www.clearpeaks.com/blog/big-data/big-data-ecosystem-spark-and-tableau

"Horses for Courses"

- Conventional RDBMS will continue play an important and significant role in OLTP (Online Transactions Processing)
- Increasingly now a *range* of database products are available, need to select appropriate product/model for task at hand.

FIT2094 Exam

2017 Exam Format

- **2 HOUR writing**
- **30 minutes reading and noting**
- 100 marks 50% of your final mark in FIT2094.
  - Minimum to pass FIT2094 overall:
    - 40% non-exam, 40% exam and 50% overall
- Questions:
  - Part A: 10 multiple choice questions (10 marks)
  - Part B: 5 questions – theory and application (90 marks)
  - Sample questions on Moodle.

Week 2 – Relational Model

- Relational model properties.
- Keys
  - Superkey, Candidate Key, Primary Key
  - Foreign Key
- Data Integrity
  - Entity integrity
  - Referential Integrity

Week 3 - 4 – Data Modelling

- Conceptual vs Logical Level
- Entity
  - Strong vs weak
  - Associative entity
- Multivalued attributes
- Relationship
  - Type : one-to-one, one-to-many, many-to-many
  - Cardinality and Participation
  - Identifying vs Non-identifying.
- Mapping from Conceptual to Logical
  - E.g. Mapping many-to-many

Week 5 – Normalisation

- UNF to 3 NF
  - UNF to 1 NF – remove repeating group.
  - 1NF to 2 NF – remove partial dependency.
  - 2NF to 3NF – remove transitive dependency.
- Dependency diagrams
- Be careful in choosing the PK!

Week 6 – Data Definition Language

- CREATE TABLE statements
  - Primary key definition
  - Foreign key definition
  - Other Constraints
- INSERT
  - Adherence to referential integrity constraints
    - Order of insertion
- Oracle Sequence
- UPDATE (DML)
- DELETE (DML)

Week 7, 9 and 10 – SQL

- Single table retrieval with predicate
- Join
  - Natural join
  - Outer join
- Aggregate functions
- Set Operators
- Subquery
- Oracle functions

Week 8 – Transaction Management

- Transaction.
- ACID properties.
- Transaction problems.
- Transaction management with locks.
- Restart and Recovery using Transaction Log.

Week 11 – PL/SQL

- Web database connectivity
  - Basic understanding of
    - Database middleware
    - Web to database middleware
    - Using PHP to communicate with databases

http://blog.proqc.com/administrative-professionals-quality-thank-you/