MONASH University

MONASH
INFORMATION
TECHNOLOGY

Week 12 -
Database Current/Future Trends
Exam Preparation

FIT2094 - FIT3171 Databases
Clayton Campus S1 2019.

GROUP OF EIGHT AUSTRALIA

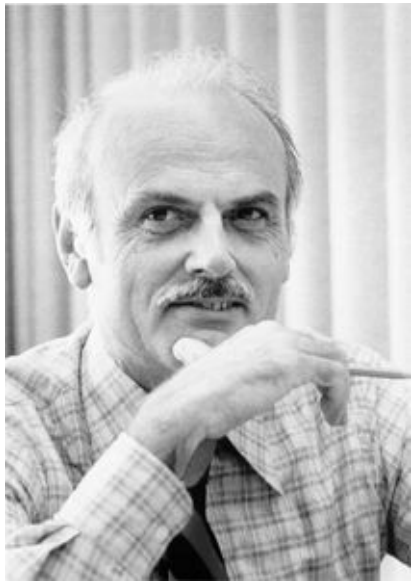
Overview

- **Hour 1**
  - –Database current and future trends
  - –Database-related industry skills/trends (ca. 2019)

… then COFFEE BREAK!

- **Hour 2**
  - –Exam preparation
  - –And important warnings

Database Hall of Fame



E.F “Ted” Codd
- you know him

Larry Ellison
- Oracle

Peter Chen
- you know him

Michael Stonebraker
- Postgres
- Turing Award
- SciDB
(refer: Wikipedia)

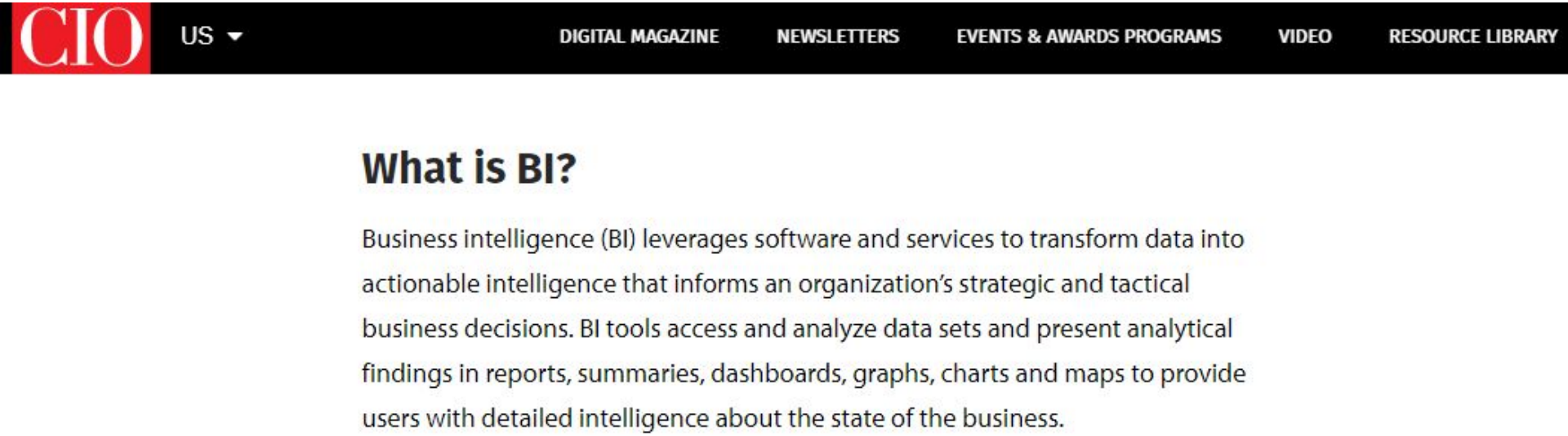
Business Intelligence and Decision Support

Img src: @adeolueletu at Unsplash

What is BI?

- CIO - Pratt (2017)

  https://www.cio.com/article/2439504/business-intelligence-definition-and-solutions.html



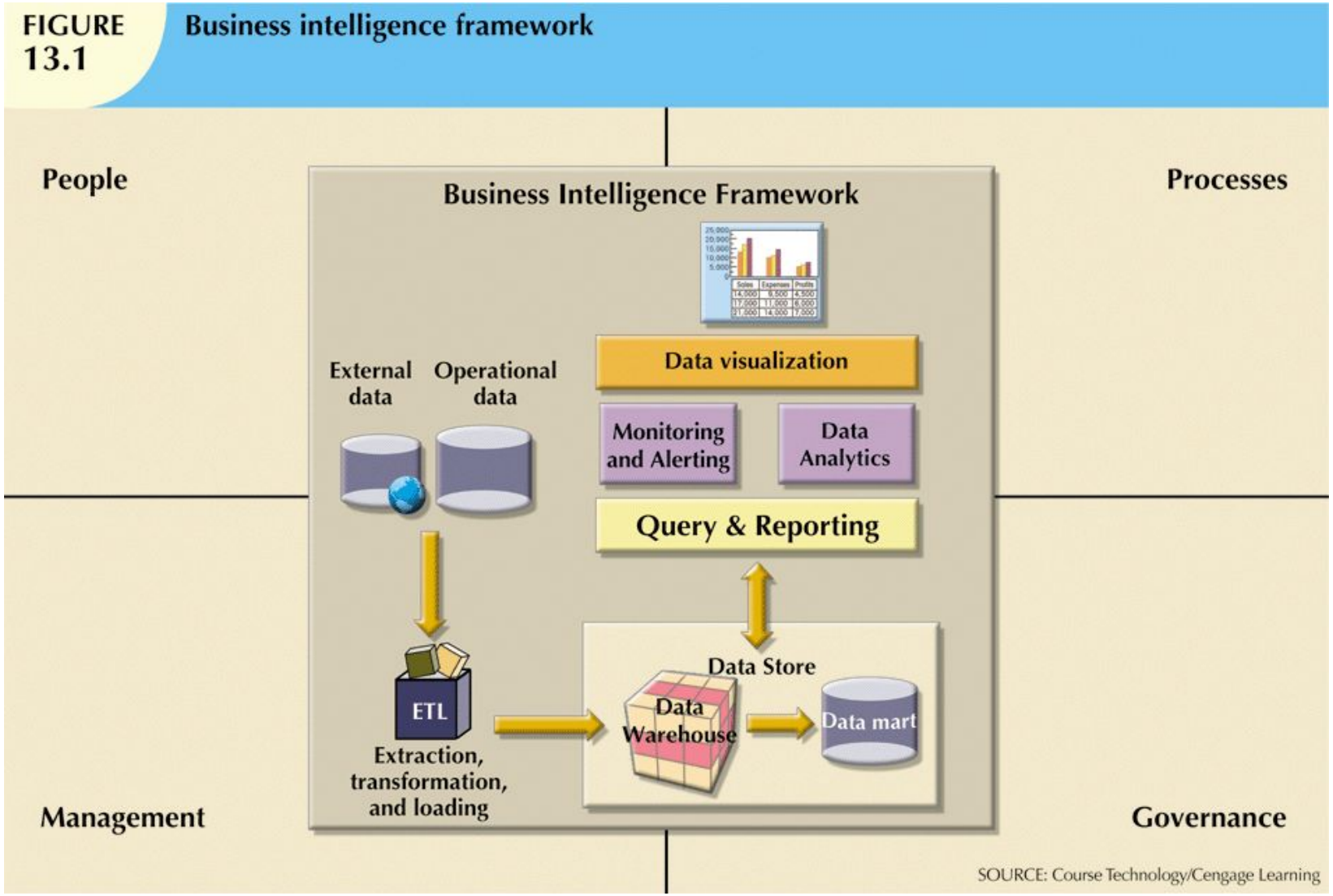


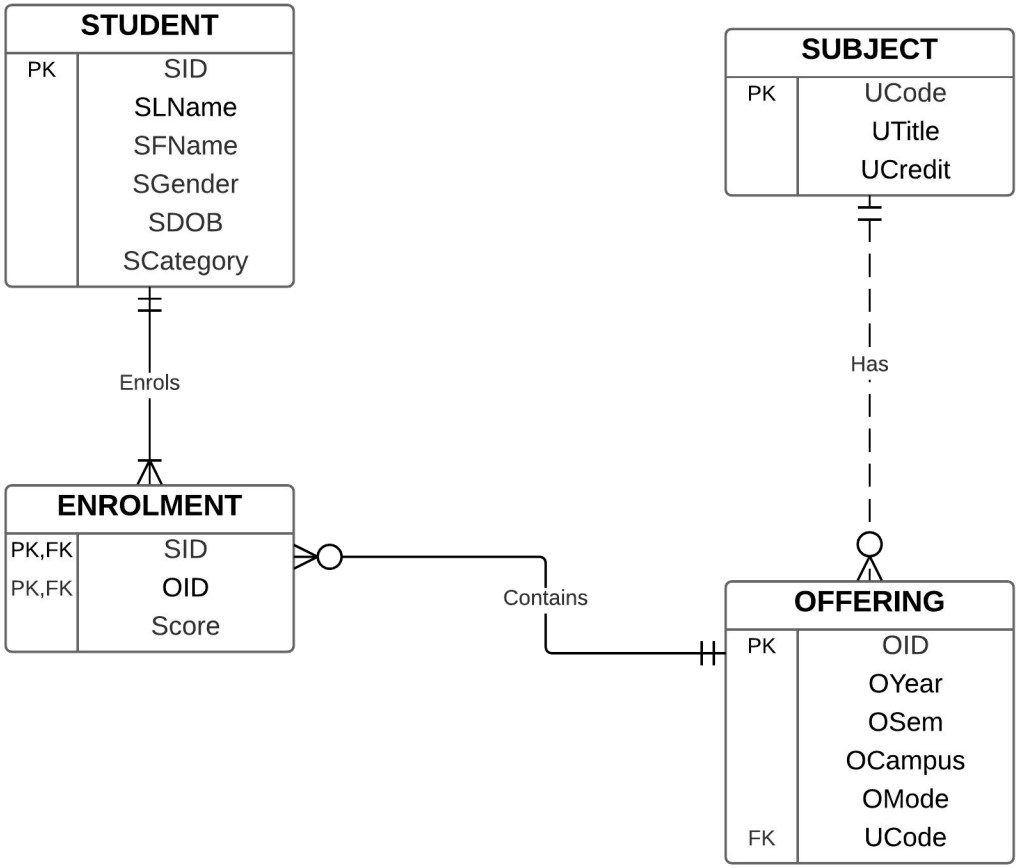
FIGURE 13.1
Business intelligence framework
People
Processes
Business Intelligence Framework
External data
Operational data
Data visualization
Monitoring and Alerting
Data Analytics
Query & Reporting
ETL
Extraction, transformation, and loading
Data Store
Data Warehouse
Data mart
Management
Governance
SOURCE: Course Technology/Cengage Learning

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 Database



[Clayton Q&A] - Why is this considered Operational level?
How can it be used for Strategic/Tactical purposes?

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.

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 (ETL)
  - 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

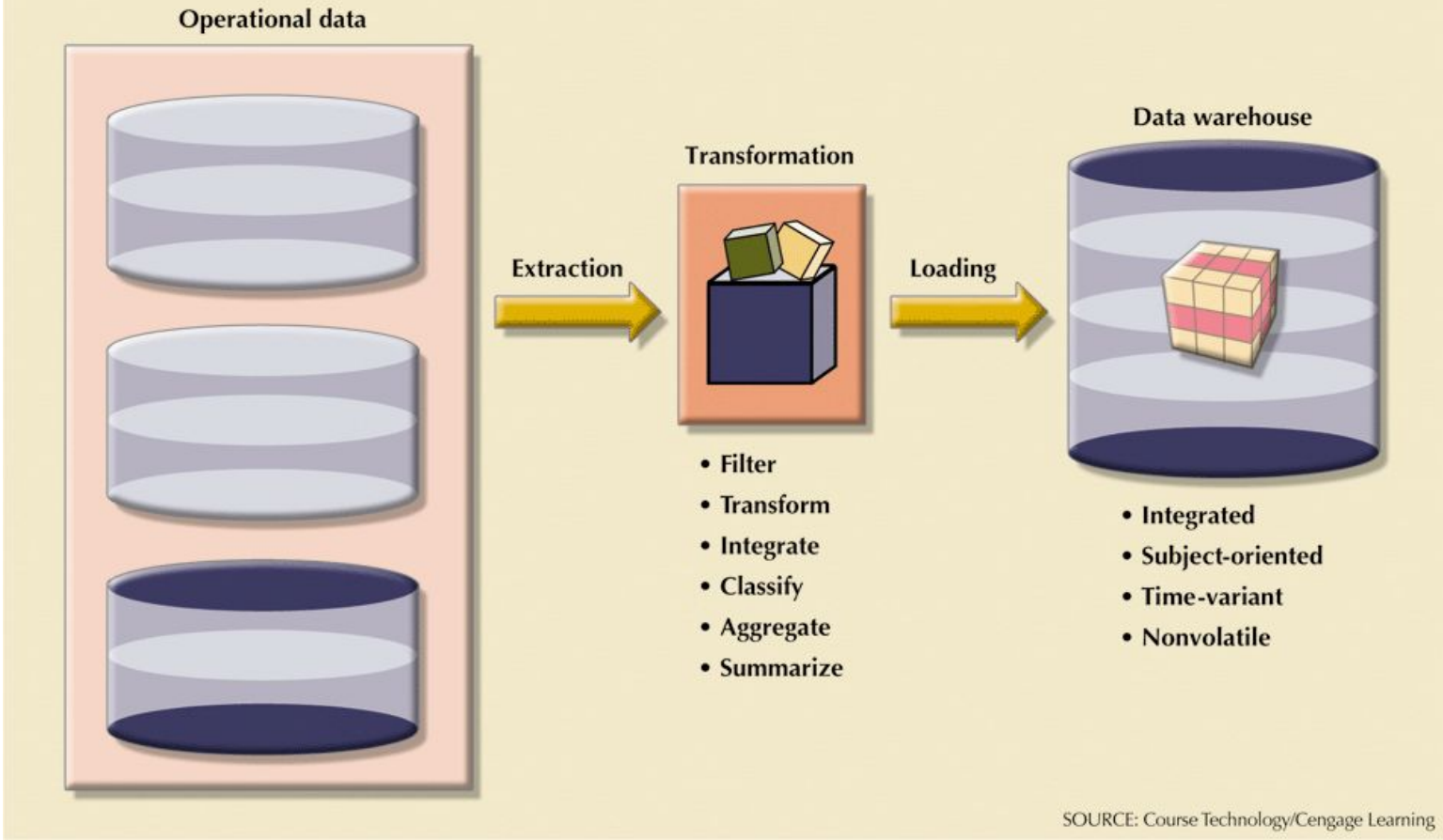
Tip of the iceberg -
More to learn in DW / Advanced DB units

Img src: @carlheyerdahl at Unsplash

The Data Warehouse (more in FIT3003)

- Database size
  - In 2013, eBay had around 90 Petabytes of data in its data warehouses (90,000 Terabytes) https://www.itnews.com.au/news/inside-ebays-90pb-data-warehouse-342615
    - "three systems, with about 7.5PB in a Teradata enterprise data warehouse, 40PB on commodity Hadoop clusters and 40PB on 'Singularity': a custom system for performing deep-dive analysis on semi-structured and relational data." - ITNews
  - "SAP ... Guinness World's record for largest data warehouse at 12.1 petabytes (PB)." https://blogs.saphana.com/2014/03/05/guinness-world-record-largest-data-warehouse/
  - 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

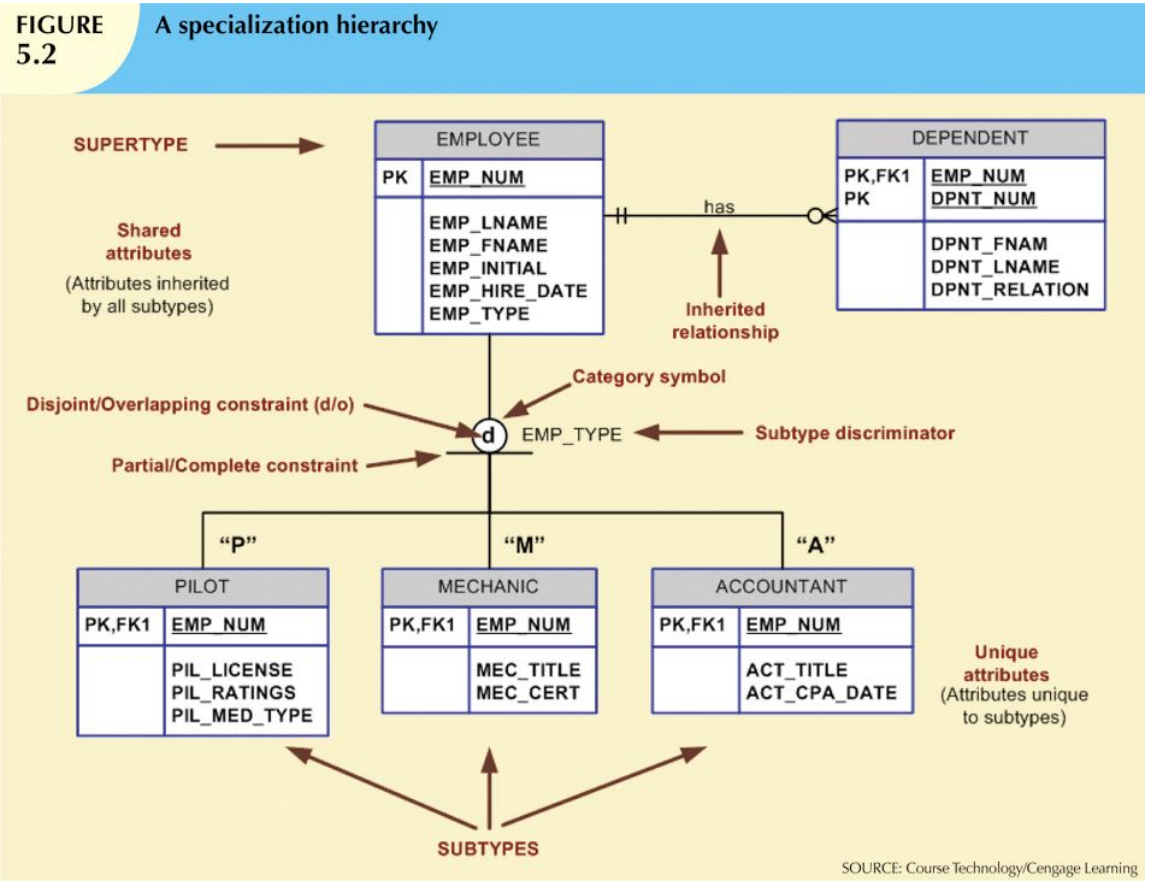


SOURCE: Course Technology/Cengage Learning

Advanced DB Design (more in FIT3176)

Logical model needs more depth - similar concept to superclass/subclass in OO

Specialization/ Generalization



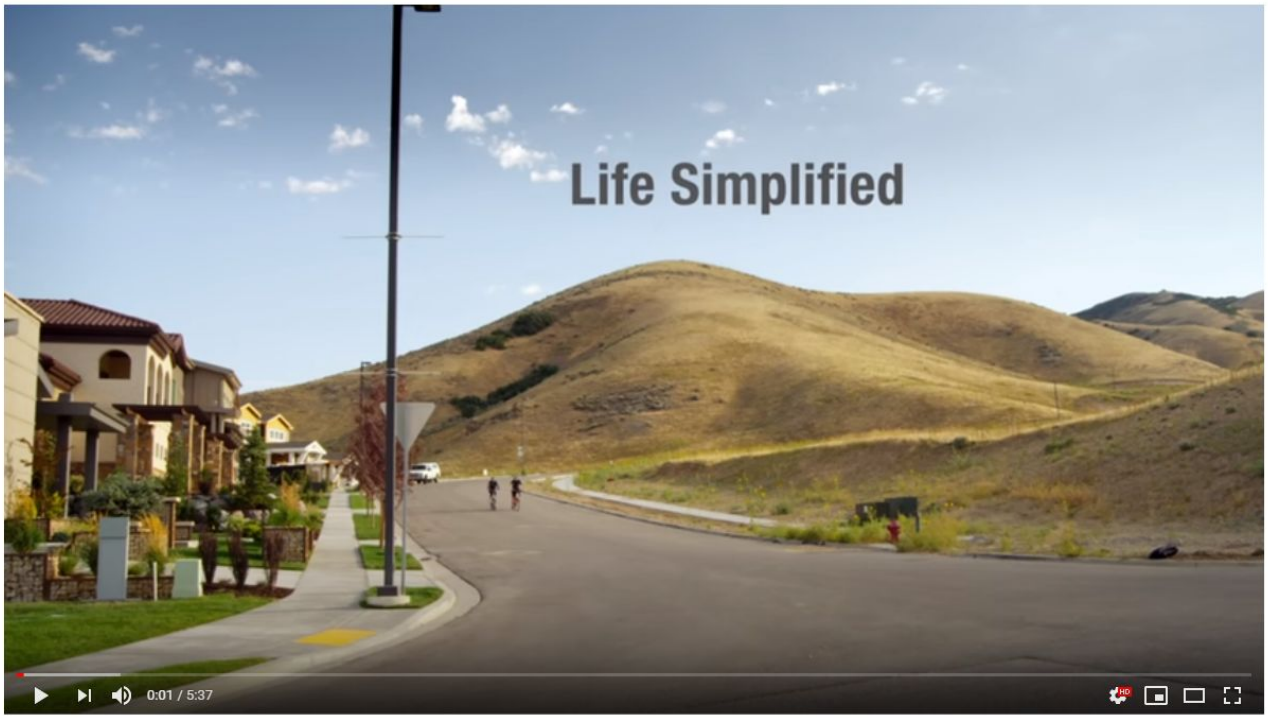
Advanced DB Design (FIT3176) continued

- E/R Diagram/Logical model is not complete enough
- Advanced SQL and PL/SQL
  - Triggers, Procedures and Packages
- XML
- How fast is running a query?

Big Data
(with case study - IoT)

Img src: @franki at Unsplash

[Internet of Things (IoT)
https://www.youtube.com/watch?v=NjYTzvAVozo](https://www.youtube.com/watch?v=NjYTzvAVozo)

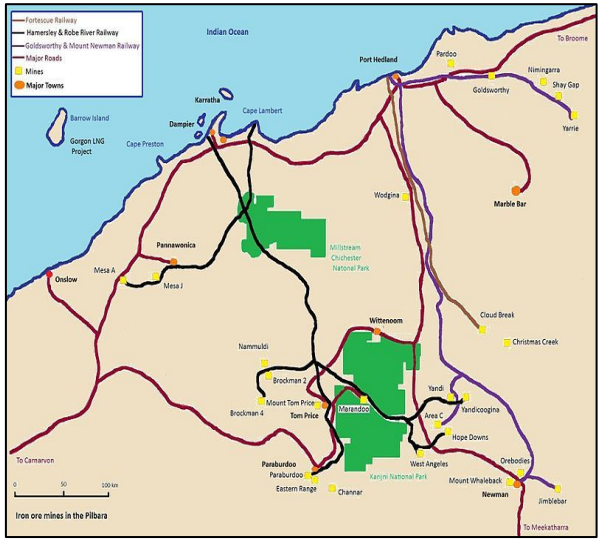


What is happening with data?

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

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

Case Study: Railway In Mining (Credit to Lindsay and team)



- 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

Case Study: Railway In Mining (Credit to Lindsay and team)

Challenges

(1)
Expensive sensors that require professionals to maintain

(2)
Large volume of data generated by the sensors

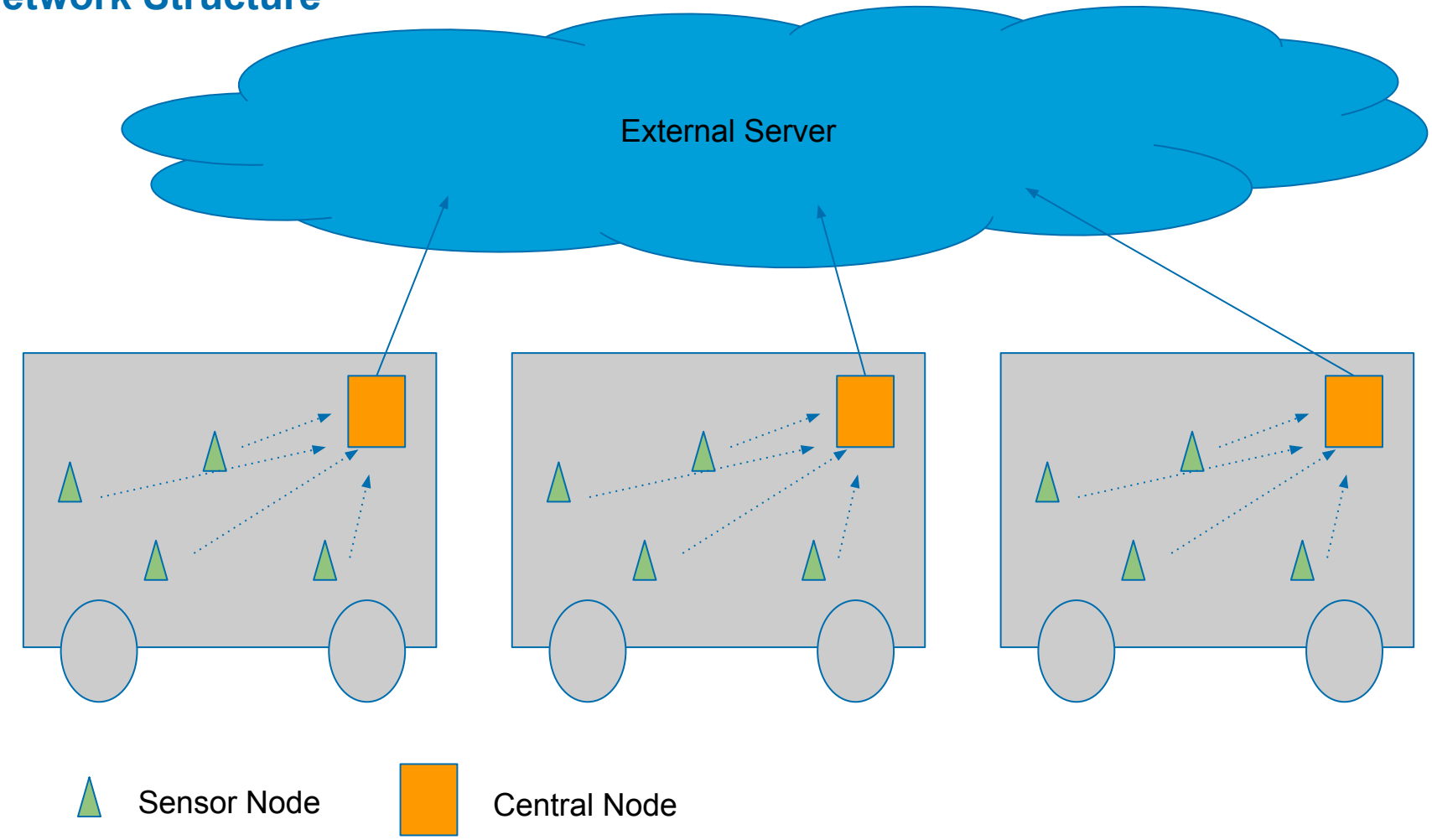
→ **Cheap, self-configured, massive array of sensors**

→ **Fast data processing and retrieval**

Needs expertise from Eng and IT

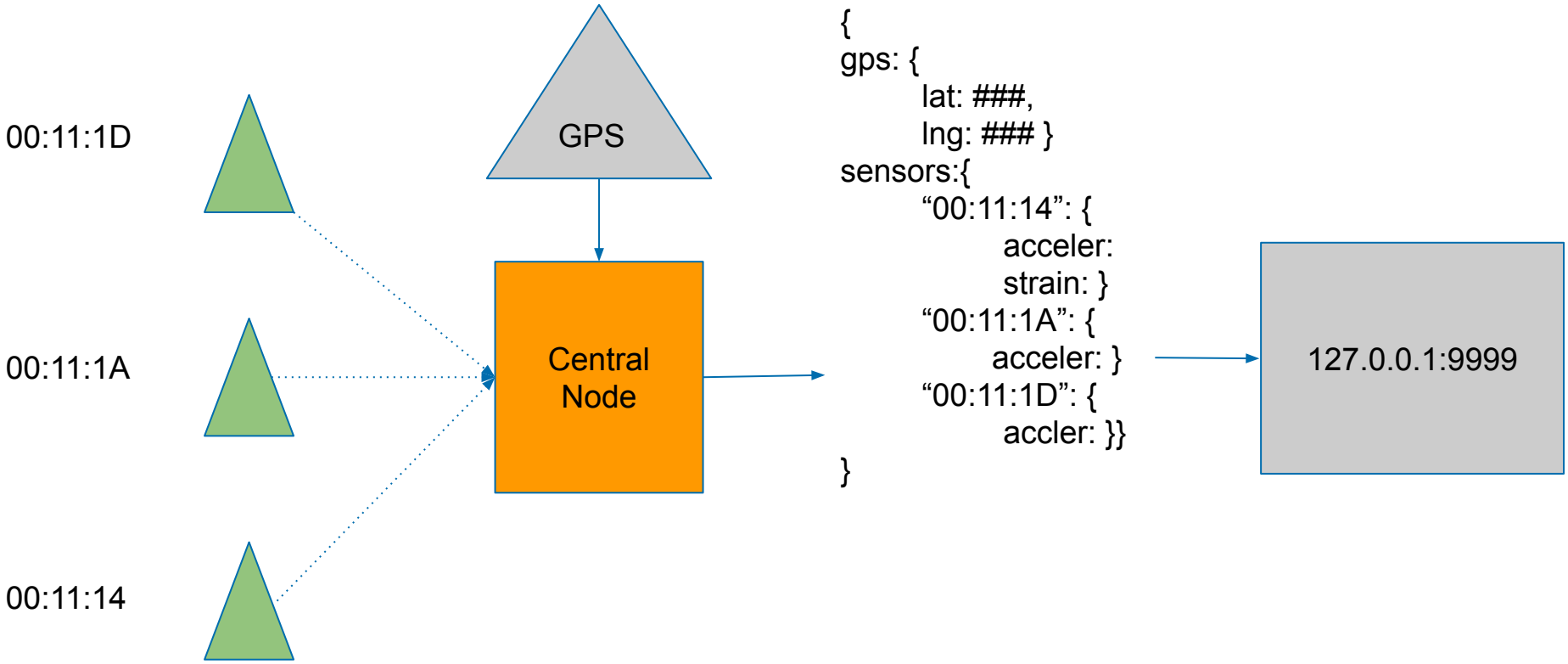
Case Study: Railway In Mining (Credit to Lindsay and team)

Network Structure



Case Study: Railway In Mining (Credit to Lindsay and team)

Central Node Process



Case Study: Railway In Mining (Credit to Lindsay and team)

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

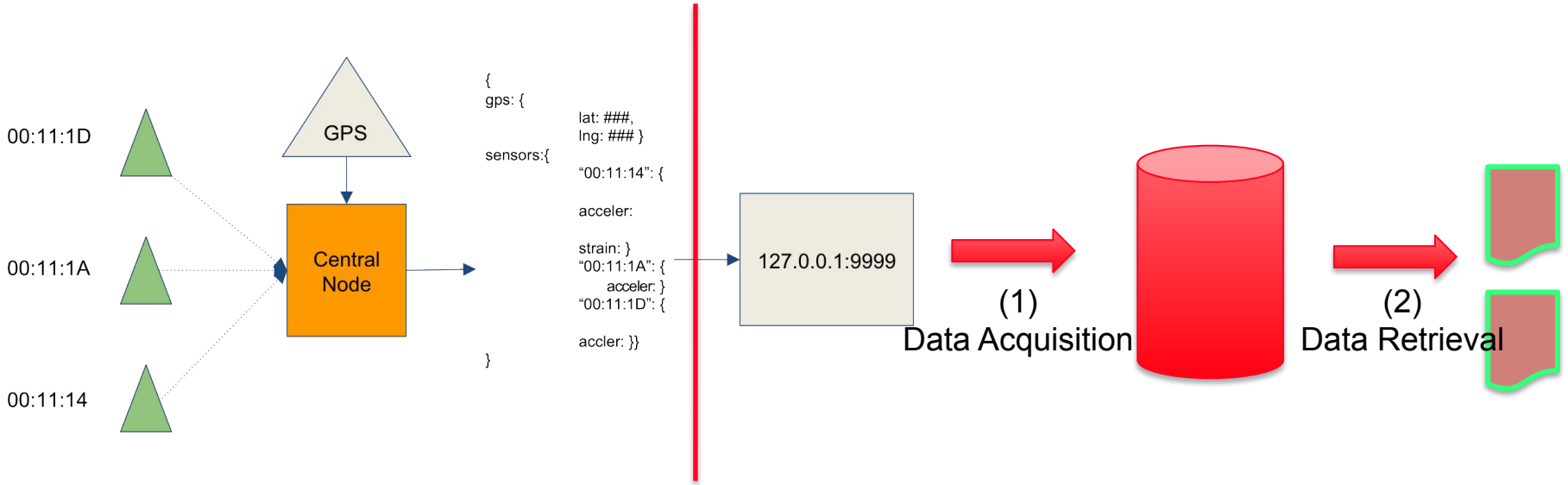
```
Welcome to Ubuntu 14.04.3 LTS (GNU/Linux 3.13.0-46-generic x86_64)

 * 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
}
```

Case Study: Railway In Mining (Credit to Lindsay and team)

Big Data Processing



Two main problems:

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

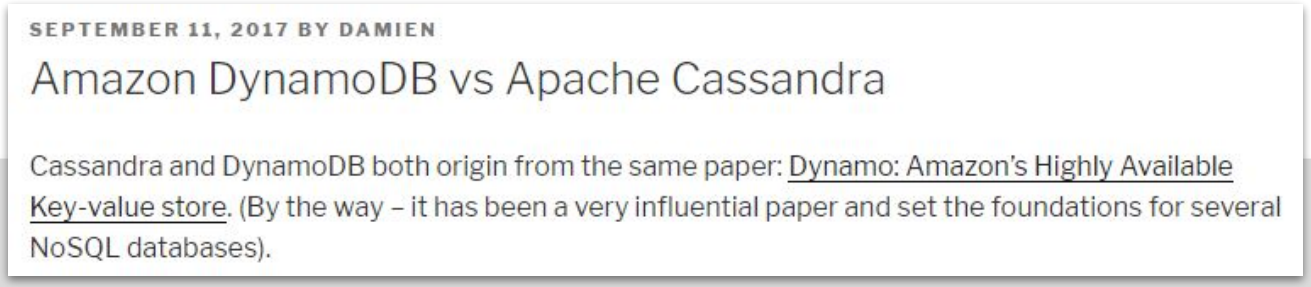
Scaling

- "Big Data" -- "**data that displays the characteristics of volume, velocity, variety (the 3 Vs)**" (Coronel & Morris, 2018)
- 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

[Clayton Q&A] - Relate the 3Vs to the case study?
Also relate the 3Vs to e.g. Facebook?

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: DynamoDB
    - http://www.read.seas.harvard.edu/~kohler/class/cs239-w08/decandia07dynamo.pdf
  - Apache Cassandra
    - http://www.beyondthelines.net/databases/dynamodb-vs-cassandra/



SEPTEMBER 11, 2017 BY DAMIEN

Amazon DynamoDB vs Apache Cassandra

Cassandra and DynamoDB both origin from the same paper: Dynamo: Amazon's Highly Available Key-value store. (By the way – it has been a very influential paper and set the foundations for several NoSQL databases).

Scaling continued

- NoSQL: “**new generation of [DBMSes] that is not based on the traditional relational database model**.” (Coronel & Morris, 2018)
  - Recommended reading: Coronel & Morris 14-3 (3a to 3d inclusive)
  - Very accessible overview.

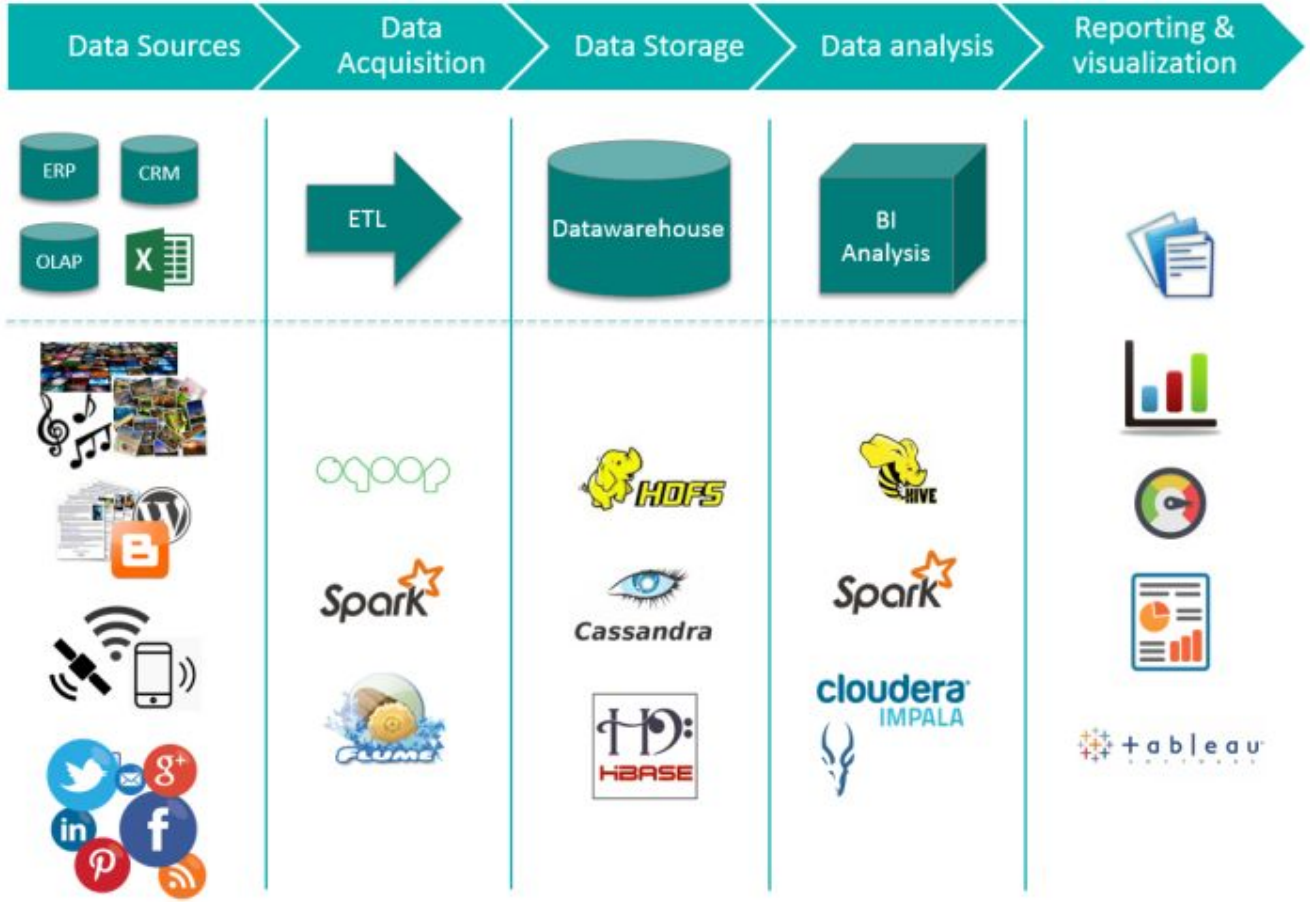
- Term "NoSQL" coined by John Oskarsson in 2009 after calling a ..."free meetup about “open source, distributed, non relational databases” or NOSQL for short"…
  - http://blog.oskarsson.nu/post/22996139456/nosql-meetup

- Characteristics
  - Non relational, mostly open source, distributed (cluster friendly), schema-less (no fixed storage schema)
  - See MongoDB "NoSQL Databases Explained"

Fast Data Processing

- 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

Lindsay anecdote: "Horses for Courses"

- “it is important to choose suitable [DBMSes]... for particular activities because every [DBMS]... has different [characteristics]”

  –Paraphrased from https://dictionary.cambridge.org

- 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.

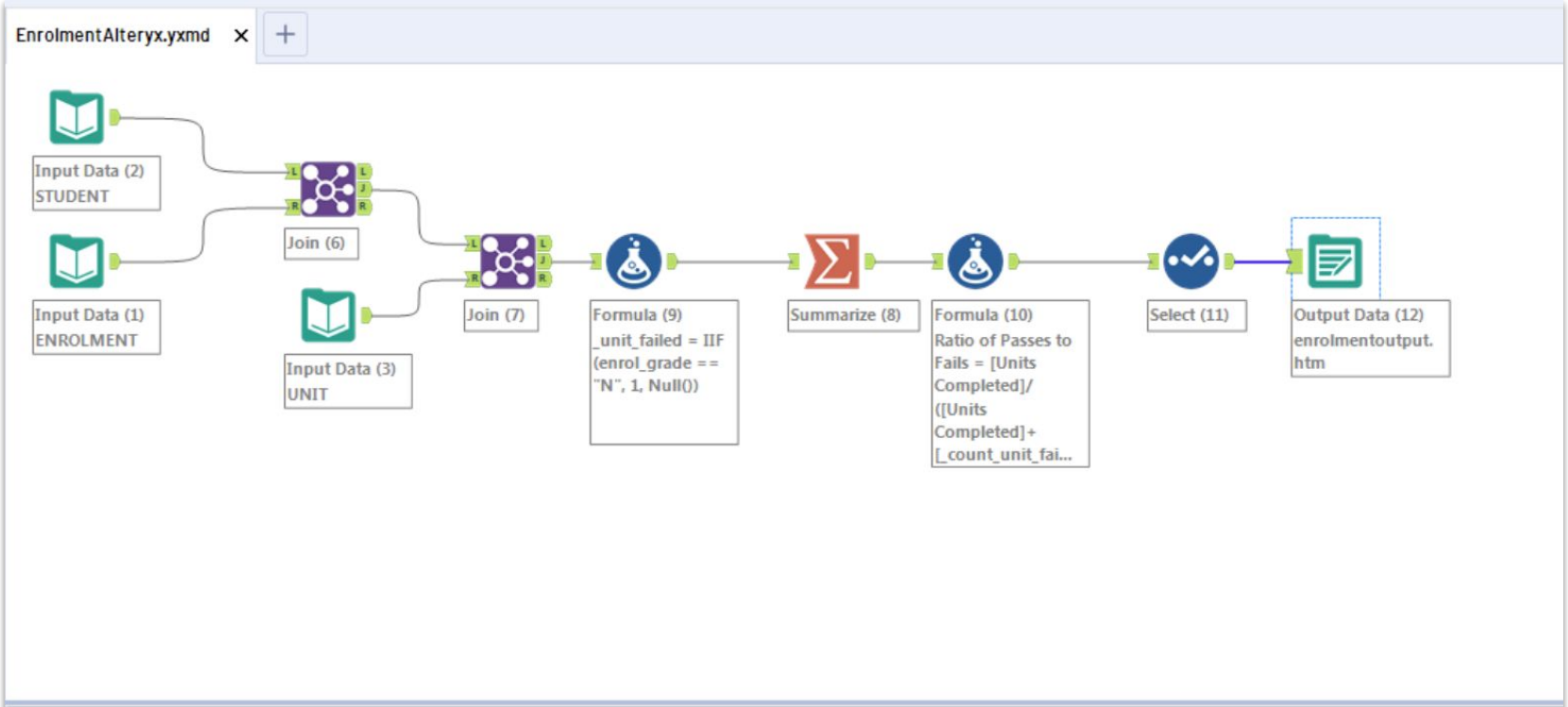
Database-related industry skills/trends (ca. 2019)

This is just a small list of key skills / applications that are currently industry- and career-relevant…
(with our familiar ENROLMENT database)

Img src: @franki at Unsplash

Alteryx - Data Analytics/Preparation/ETL

Drag-and-drop analytics platform… “self-service data analytics ... with a platform that can discover, prep, and analyze all your data, then deploy and share analytics at scale” - Alteryx.com.

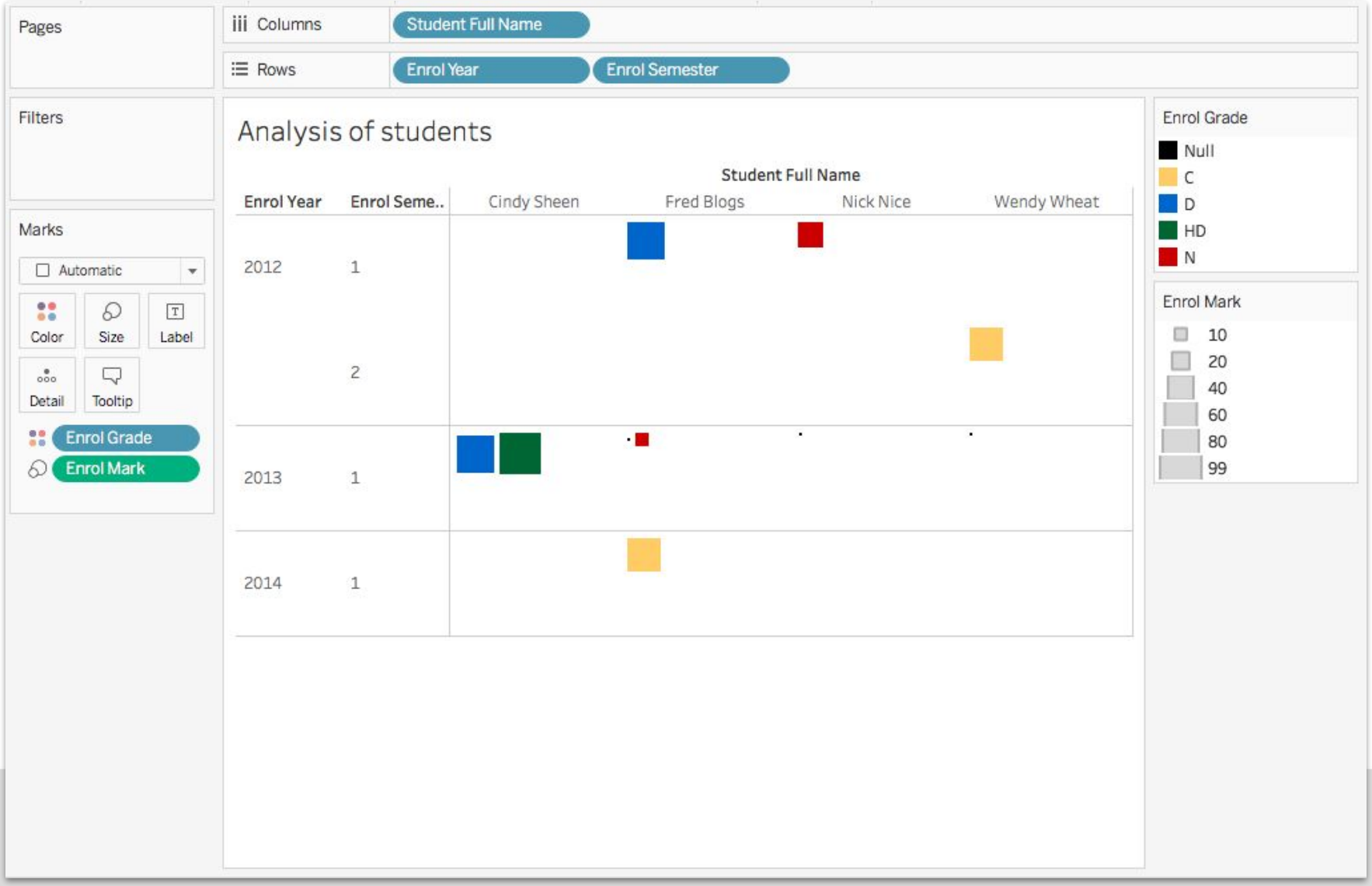


Results - Output Data (12) - Input

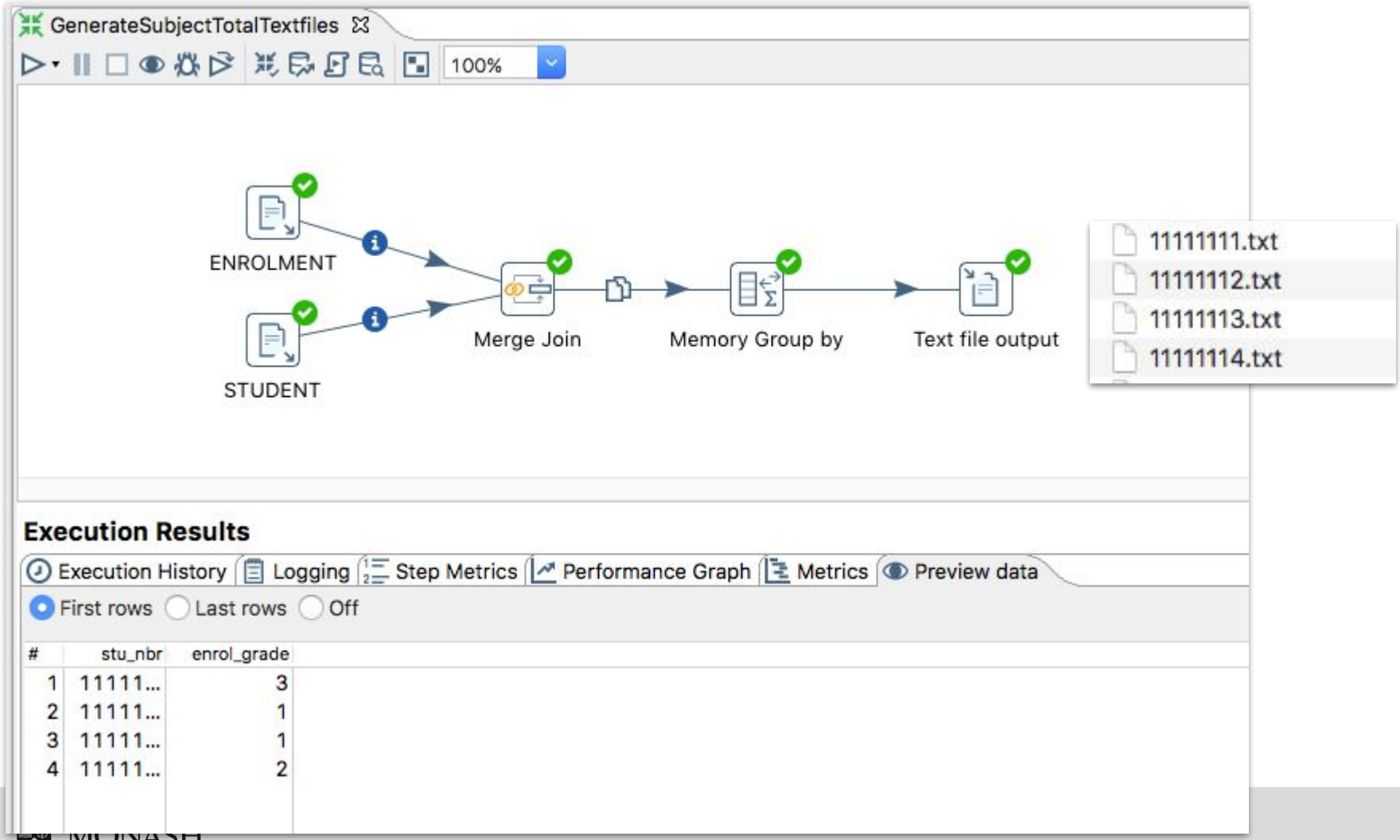
6 of 6 Fields | Cell Viewer | 4 records displayed

Record #	Student Number	Student Last Name	Student First Name	Units Completed	Units Not Completed	Ratio of Passes to Fails
1	11111111	Blogs	Fred	3	1	0.75
2	11111112	Nice	Nick	1	1	0.5
3	11111113	Wheat	Wendy	1	1	1
4	11111114	Sheen	Cindy	2	0	1

Tableau - BI/Analytics/Dashboarding

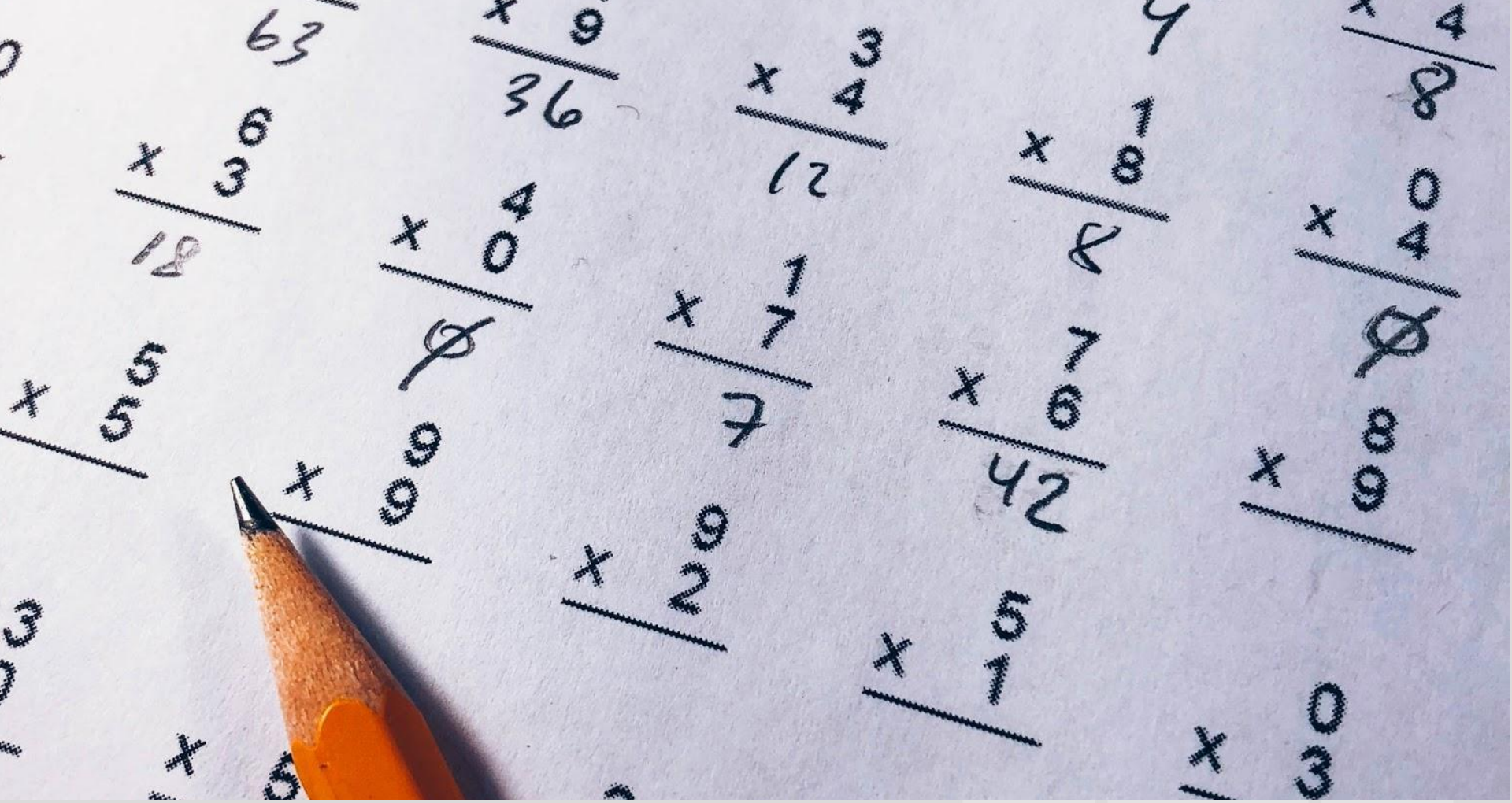


Hitachi-Vantara (Pentaho) PDI - ETL



Coffee break - see you in 10 minutes.

Img src: Jazmin Quaynor @jazminantoinette at Unsplash



FIT2094 and FIT3171 Exam Preparation

Img src: @chrisliverani at Unsplash

2019 Exam Format

- **2 HOUR writing**
- **10 minutes reading**
- 100 marks 50% of your final mark
  - BOTH FIT2094 / FIT3171
  - Minimum to pass overall:
    - 40% non-exam, 40% exam and 50% overall

- Write in exam answer script book.
  - Anything written in question sheet cannot be marked.

2019 Exam Format

- Question structure:
  - **High level marks distribution follows the sample exam.**
    - Sample exam - one for each unit code - on Moodle.
  - **5 top level Q's - please review them NOW**
    - **10m Rel Model**
    - **20m DB Design**
    - **20m Normalisation**
    - **10m Transaction Mgmt**
    - **40m 'SQL, Database Theory and Implementation' differs between FIT2094 and FIT3171 in difficulty!!!**
  - Supplementary Q's by Lindsay - please revise
    - PL/SQL, Relational Algebra, SQL
  - Reopening COPY OF prior quizzes as a compilation - NO MARKS

2019 Exam Format

- **WARNING**
  - **The next few slides are as a guide ONLY; and NOT the actual distribution/composition.**
  - **About the sample exams:**

- IMPORTANT NOTE: This Sample Exam serves to provide a general overview of the general structure of the exam paper only.

- To protect the integrity of the exam:
NO ACTUAL EXAM QUESTIONS are included; and the COMPOSITION OF THE SUBQUESTIONS are SUBJECT TO CHANGE.

- Students are reminded that all content specified by the Unit Guide is examinable, including but not limited to Pre-reading (weekly Coronel & Morris chapters) + Lecture Notes + Tute Notes + all other Moodle Material (except where explicitly stated).

Week 1-2 – Relational Model
INCLUDING BUT NOT LIMITED TO THESE TOPICS...

- Database basics
  - –Anomalies, Redundancies, etc.
- Relational model properties.
- Keys
  - –Superkey, Candidate Key, Primary Key
  - –Foreign Key
- Data Integrity
  - –Entity integrity
  - –Referential Integrity
- Relational Algebra
  - –Understanding of efficiency of solution

Week 3 - 4 – Data Modelling
INCLUDING BUT NOT LIMITED TO THESE TOPICS...

- Conceptual vs Logical Level
  - –Chen/Crows feet & UML Conceptual Modelling
- 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
INCLUDING BUT NOT LIMITED TO THESE TOPICS...

- UNF to 3NF
  - UNF to 1NF – remove repeating group.
  - 1NF to 2NF – remove partial dependency.
  - 2NF to 3NF – remove transitive dependency.
- Dependency diagrams
- Be careful in choosing the PK!
- Mapping a set of 3NF relations to a logical model

Week 6 – Data Definition Language
INCLUDING BUT NOT LIMITED TO THESE TOPICS...

- 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
INCLUDING BUT NOT LIMITED TO THESE TOPICS...

- SQL Queries in the exam
- Single table retrieval with predicate
- Join
  - Natural join
  - Outer join
- Aggregate functions
- Set Operators
- Subquery
- Oracle functions
- Triggers

Week 8 – Transaction Management
INCLUDING BUT NOT LIMITED TO THESE TOPICS...

- Transaction.
- ACID properties.
- Transaction problems.
- Transaction management with locks.
  - Deadlock detection - tables, wait for graphs, and handling
- Restart and Recovery using Transaction Log.

Week 11 – Web Database
INCLUDING BUT NOT LIMITED TO THESE TOPICS...

- Web database connectivity
  - –Understanding of the principles and ALL core concepts:
    - •Database middleware
    - •Web to database middleware
    - •Using PHP to communicate with databases
  - **–No requirement to code PHP in exam**
  - –Database design frameworks
    - •modern frameworks
    - •ORM
    - •Security

Consultations for Final Exam

- TBA Week 13
- Check Moodle under consultations
- Make use of forums
  - NB: due to heavy email volumes and forum volumes, due to e.g. A2, **your patience will be appreciated.**

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