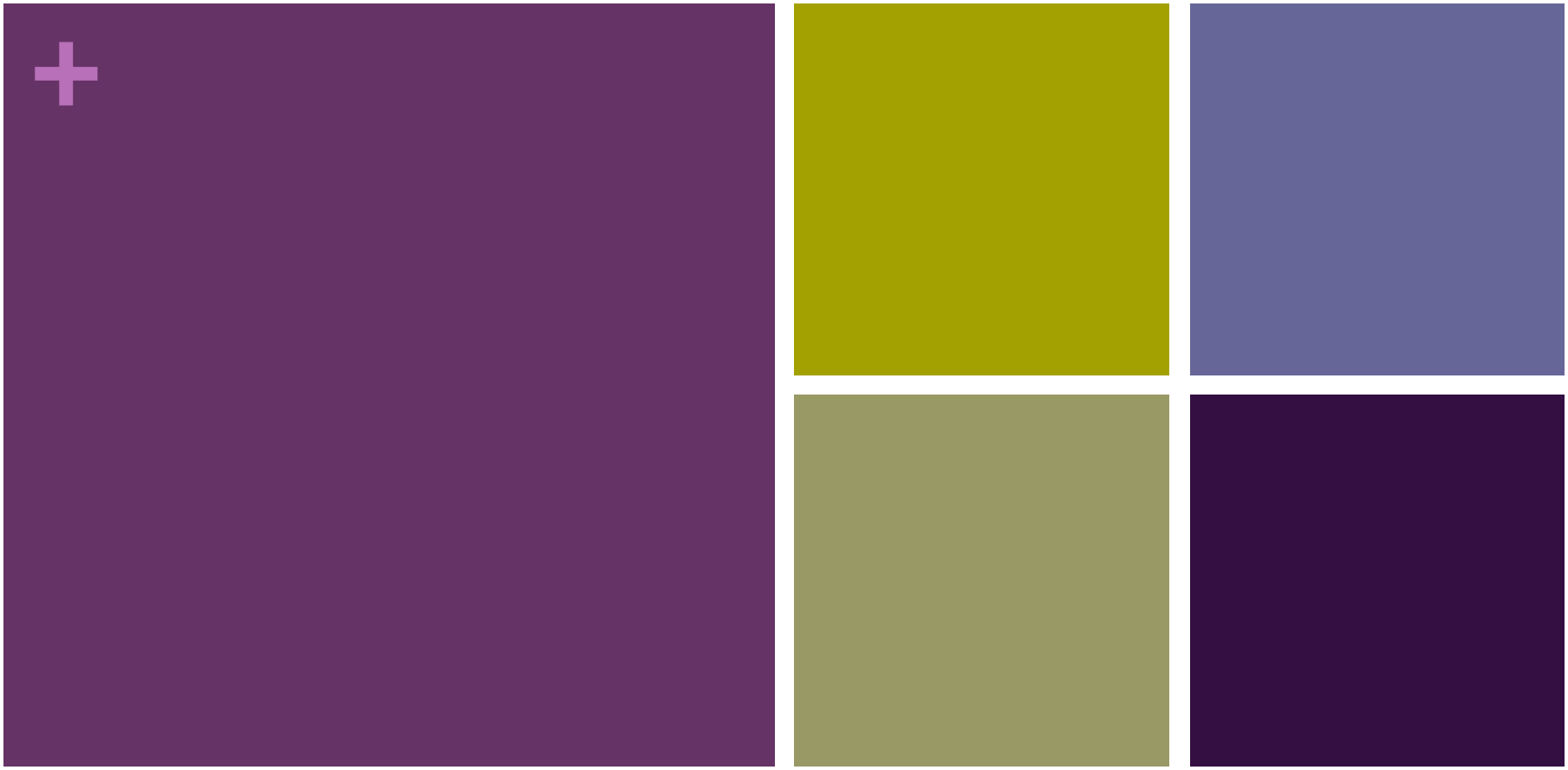
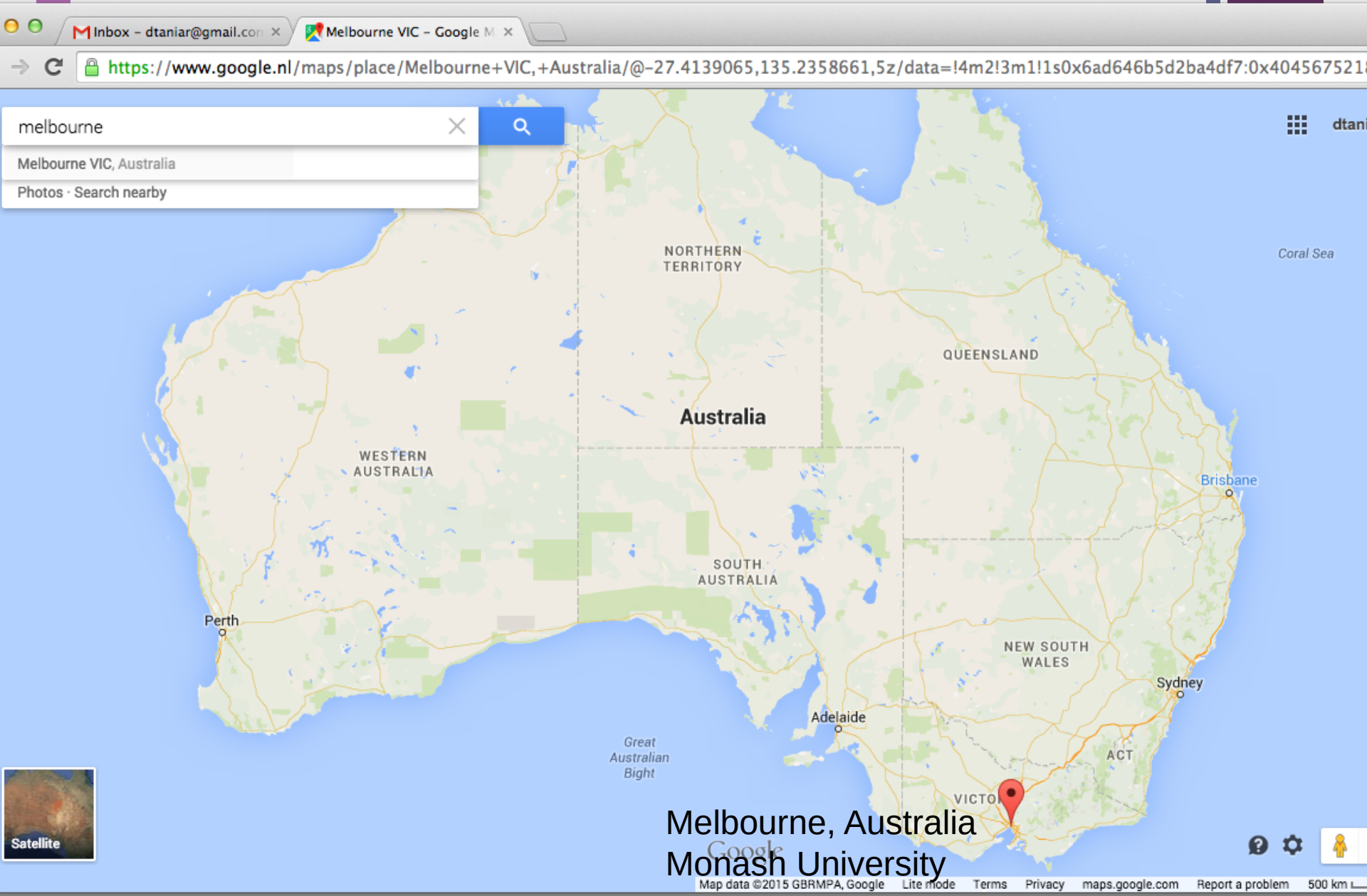


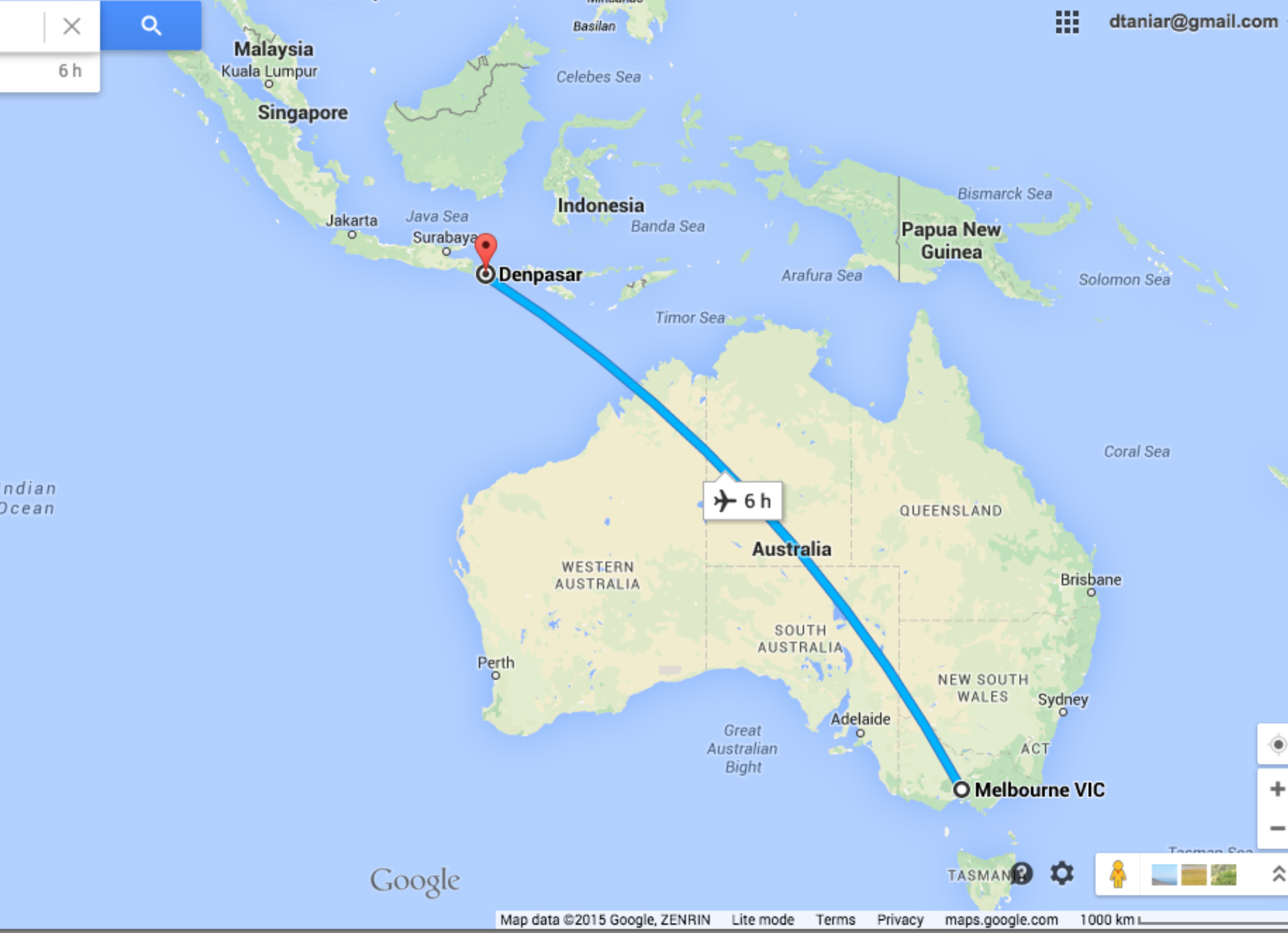
Big Data is all about data that we don’t have

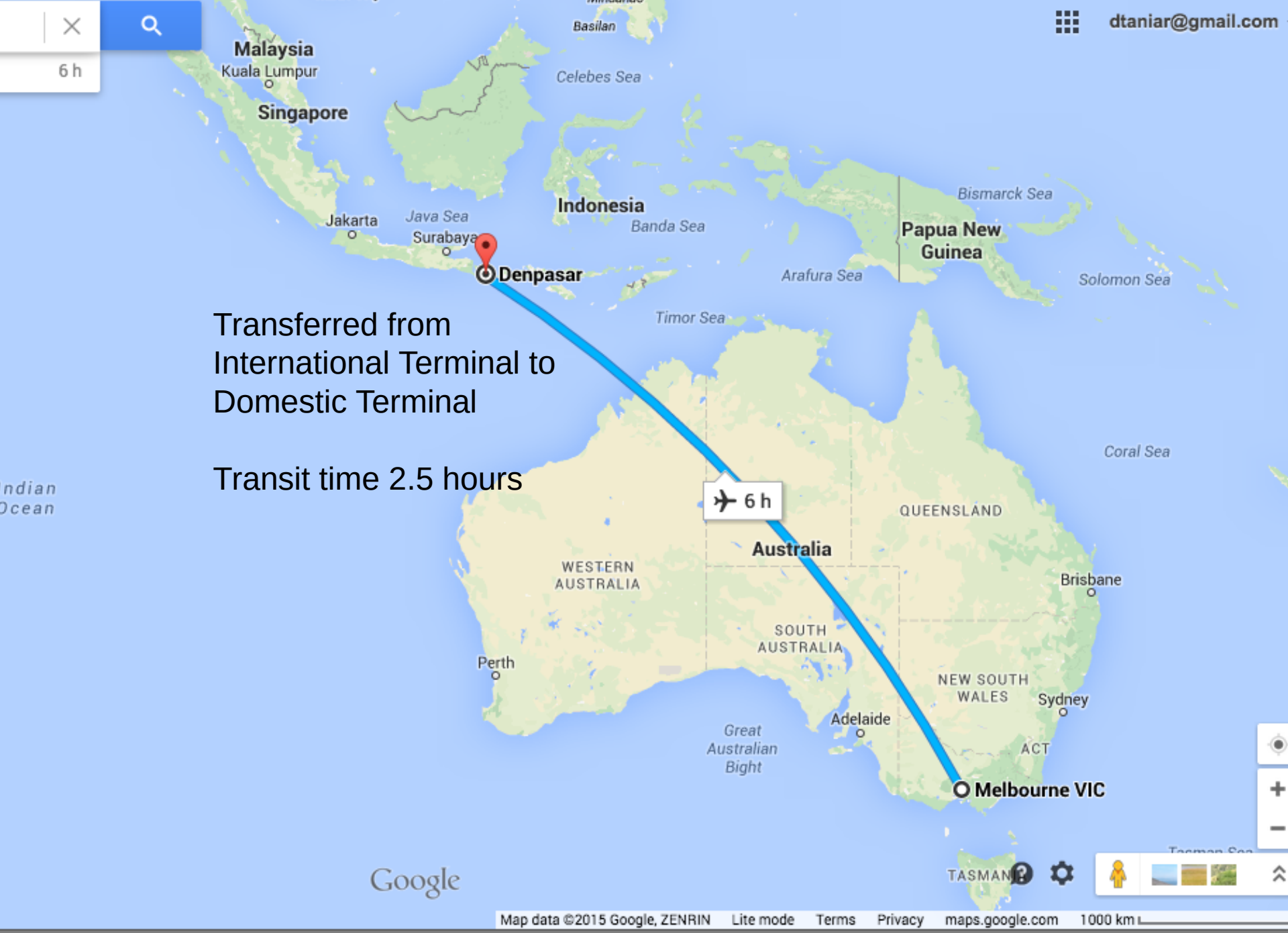
David Taniar
Monash University, Australia

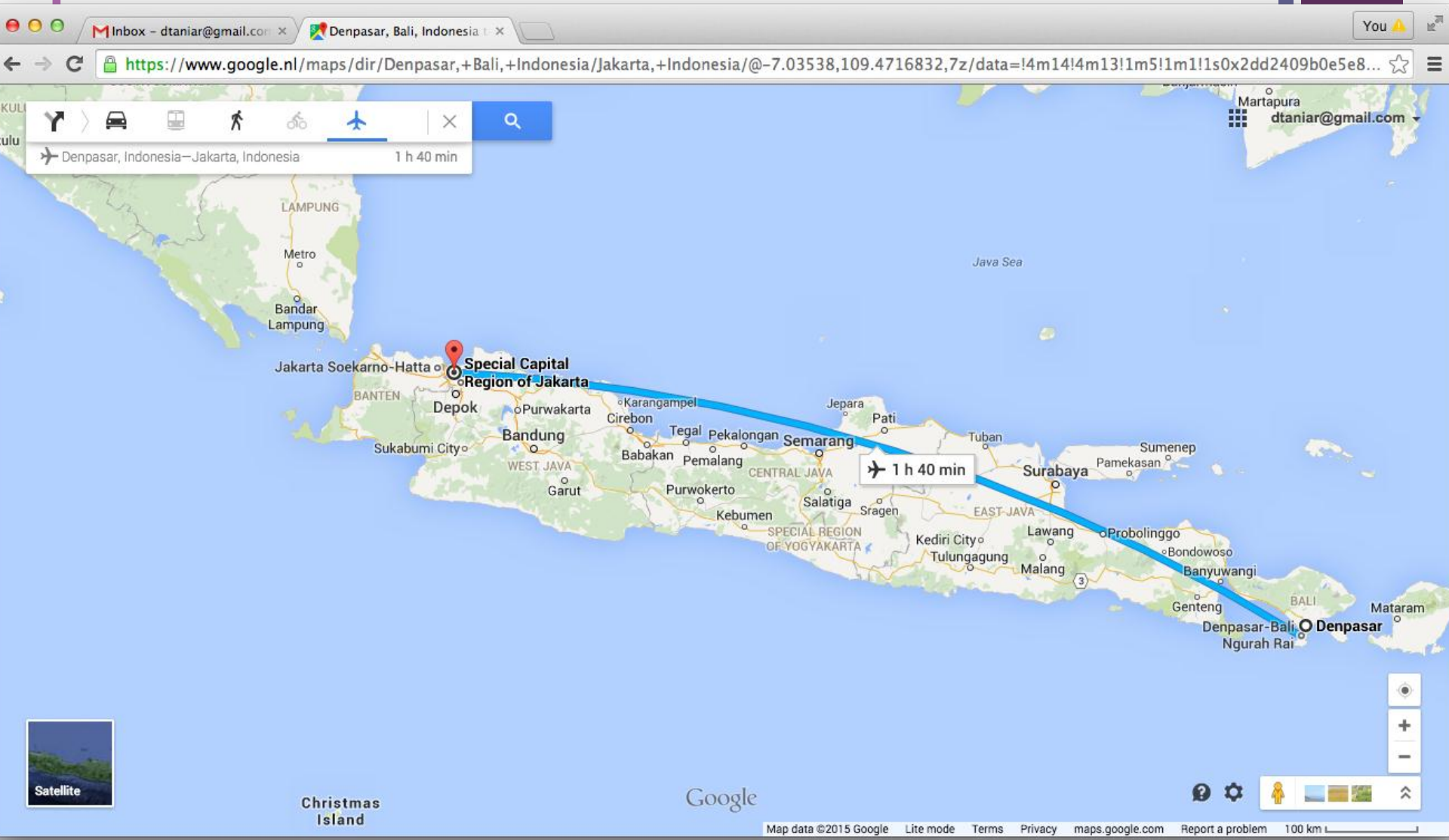
EGC’2015, Luxembourg, 28 January 2015

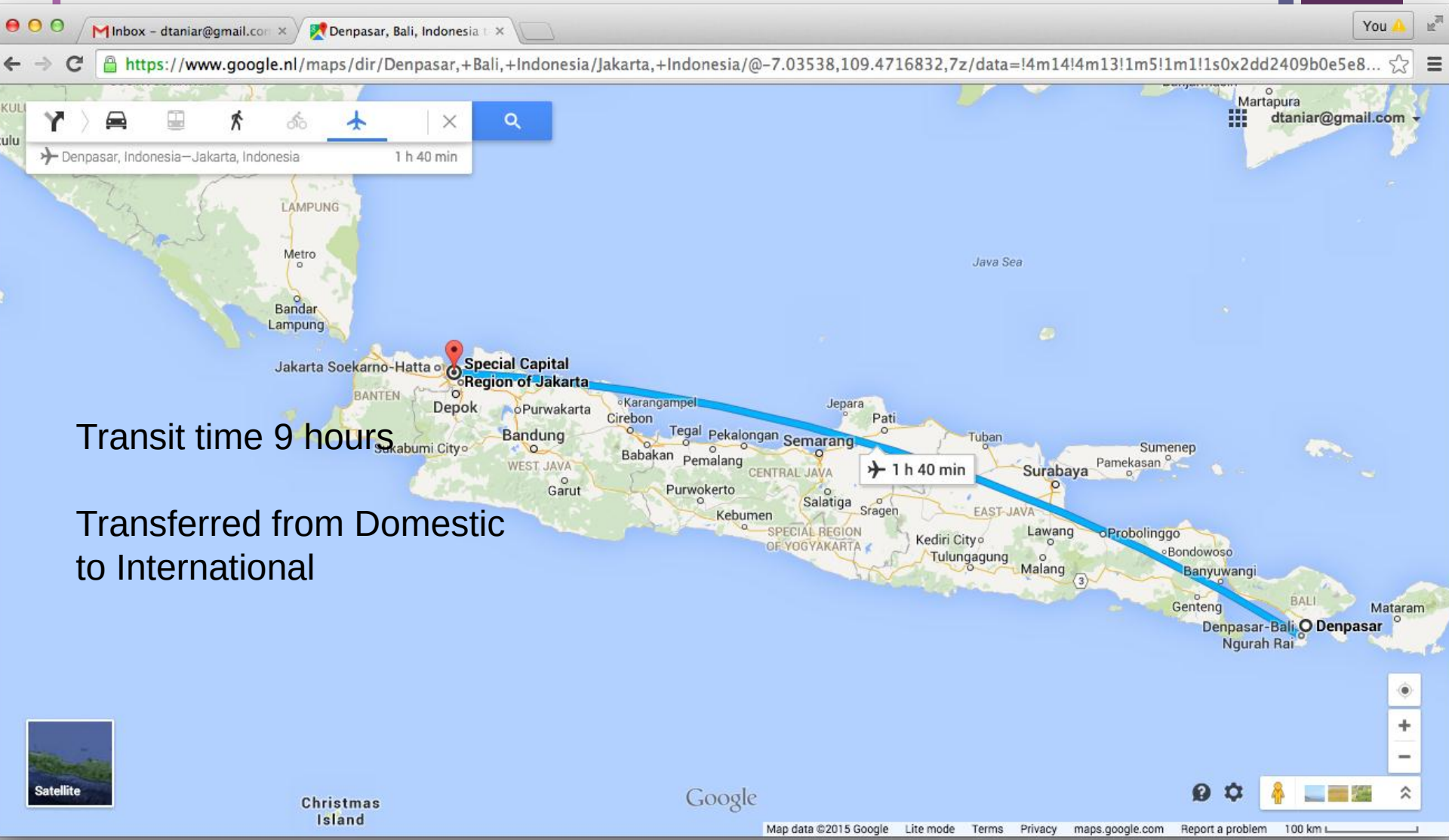


Melbourne, Australia
Monash University


dtaniar@gmail.com
6 h
Malaysia
Kuala Lumpur
Singapore
Basilan
Celebes Sea
Indonesia
Jakarta
Java Sea
Surabaya
Banda Sea
Bismarck Sea
Papua New Guinea
Denpasar
Arafura Sea
Solomon Sea
Timor Sea
Coral Sea
Indian Ocean
6 h
QUEENSLAND
Australia
WESTERN AUSTRALIA
Brisbane
SOUTH AUSTRALIA
Perth
NEW SOUTH WALES
Sydney
Adelaide
ACT
Great Australian Bight
Melbourne VIC
TASMANIA
Google
Map data ©2015 Google, ZENRIN
Lite mode
Terms
Privacy
maps.google.com
1000 km

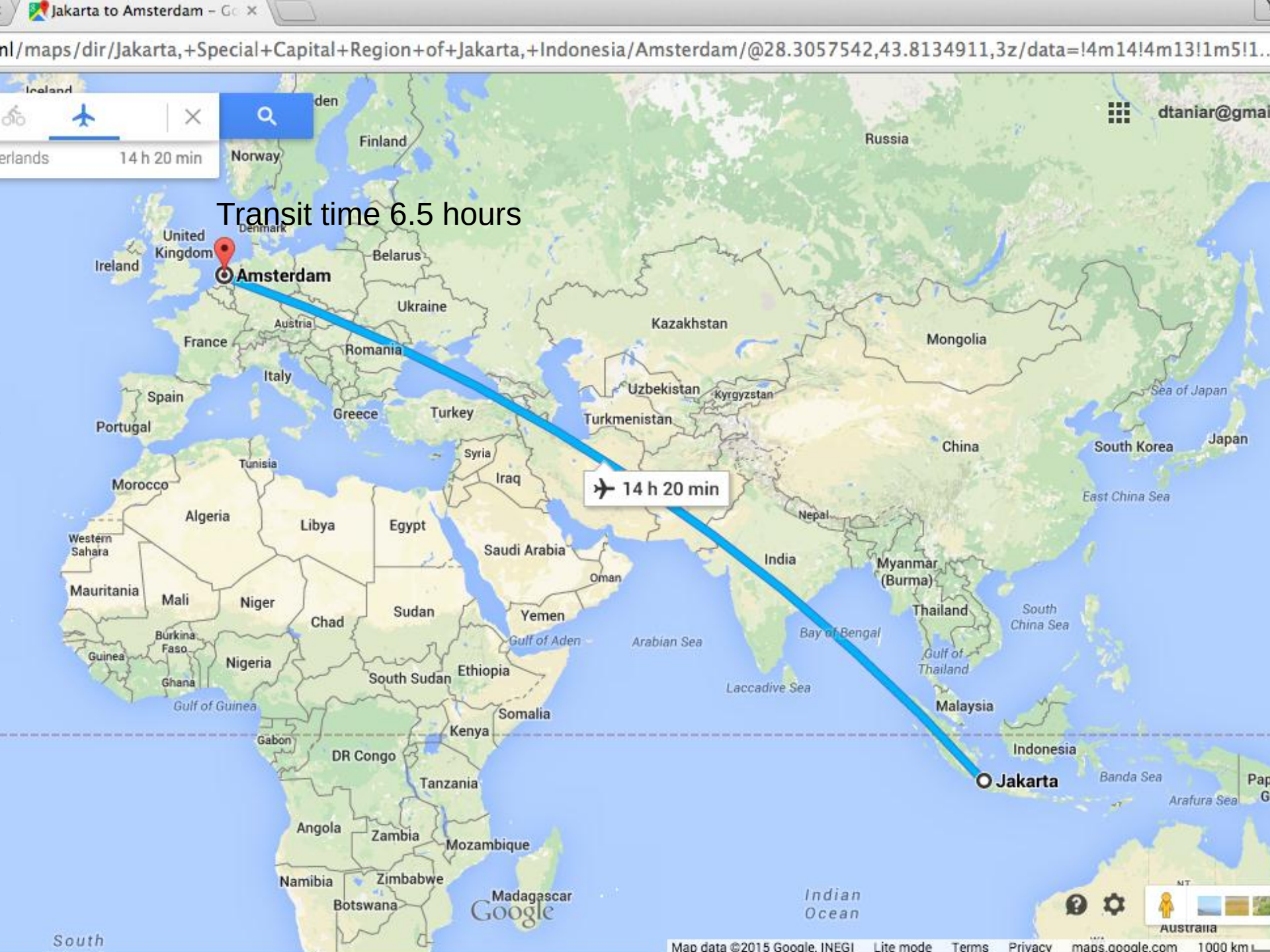

dtaniar@gmail.com
6 h
Malaysia
Kuala Lumpur
Singapore
Basilan
Celebes Sea
Indonesia
Jakarta
Java Sea
Surabaya
Banda Sea
Bismarck Sea
Papua New Guinea
Denpasar
Arafura Sea
Solomon Sea
Timor Sea
Transferred from International Terminal to Domestic Terminal
Transit time 2.5 hours
Coral Sea
6 h
QUEENSLAND
Australia
WESTERN AUSTRALIA
Brisbane
SOUTH AUSTRALIA
Perth
NEW SOUTH WALES
Sydney
Adelaide
Great Australian Bight
ACT
Melbourne VIC
TASMAN
Google
Map data ©2015 Google, ZENRIN
Lite mode
Terms
Privacy
maps.google.com
1000 km

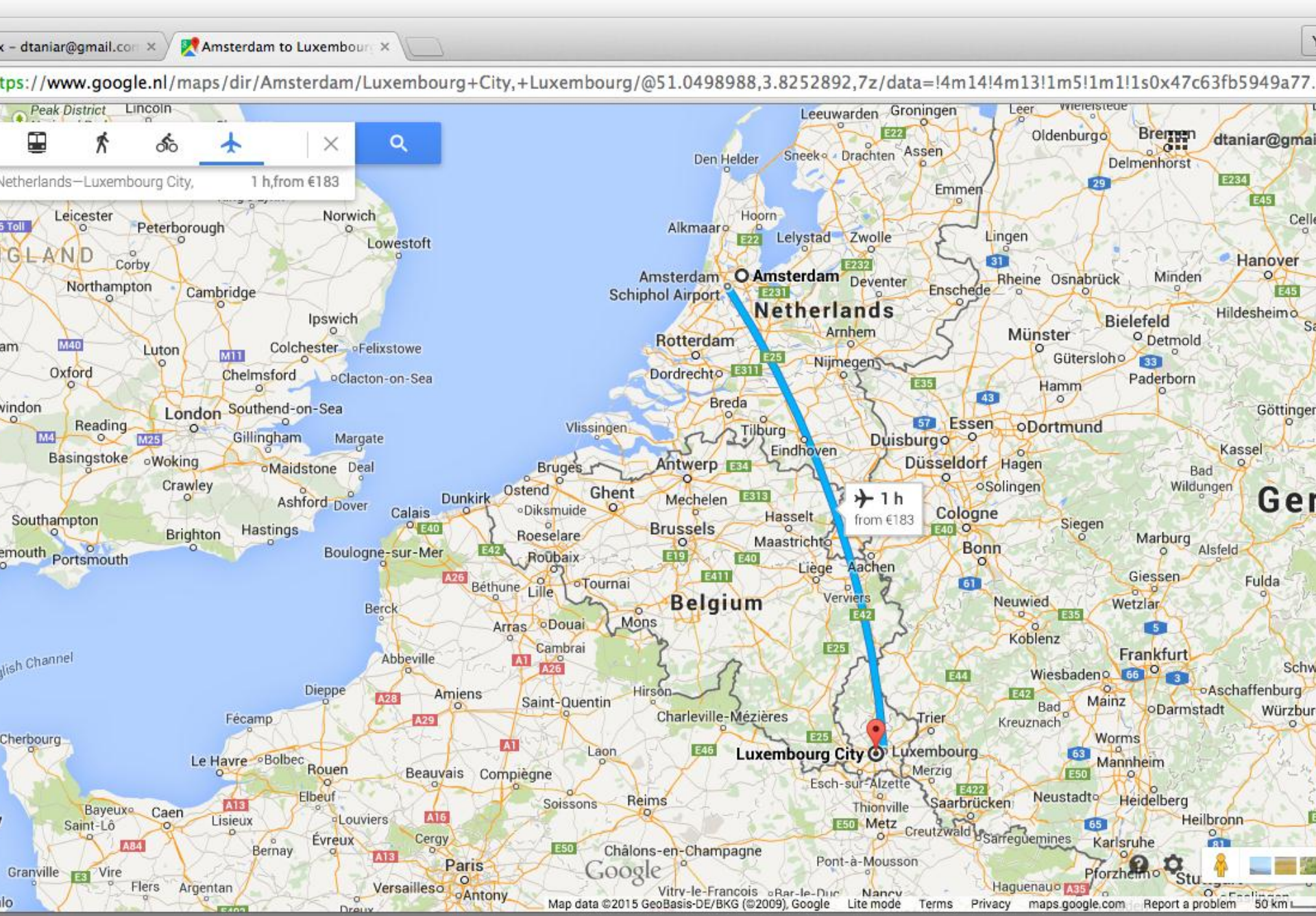

Inbox - dtaniar@gmail.com
Denpasar, Bali, Indonesia
https://www.google.nl/maps/dir/Denpasar,+Bali,+Indonesia/Jakarta,+Indonesia/@-7.03538,109.4716832,7z/data=!4m14!4m13!1m5!1m1!1s0x2dd2409b0e5e8...
Denpasar, Indonesia—Jakarta, Indonesia 1 h 40 min
dtaniar@gmail.com
Java Sea
LAMPUNG
Metro
Bandar Lampung
Jakarta Soekarno-Hatta
Special Capital Region of Jakarta
BANTEN
Depok
Purwakarta
Karangampel
Cirebon
Tegal
Pekalongan
Semarang
Jepara
Pati
Tuban
Sumenep
Pamekasan
Surabaya
Bandung
Sukabumi City
WEST JAVA
Babakan
Pemalang
CENTRAL JAVA
1 h 40 min
Garut
Purwokerto
Salatiga
Sragen
EAST JAVA
Kebumen
SPECIAL REGION OF YOGYAKARTA
Kediri City
Lawang
Probolinggo
Bondowoso
Tulungagung
Malang
Banyuwangi
Genteng
BALI
Mataram
Denpasar-Bali Ngurah Rai
Denpasar
Satellite
Christmas Island
Google
Map data ©2015 Google
Lite mode
Terms
Privacy
maps.google.com
Report a problem
100 km

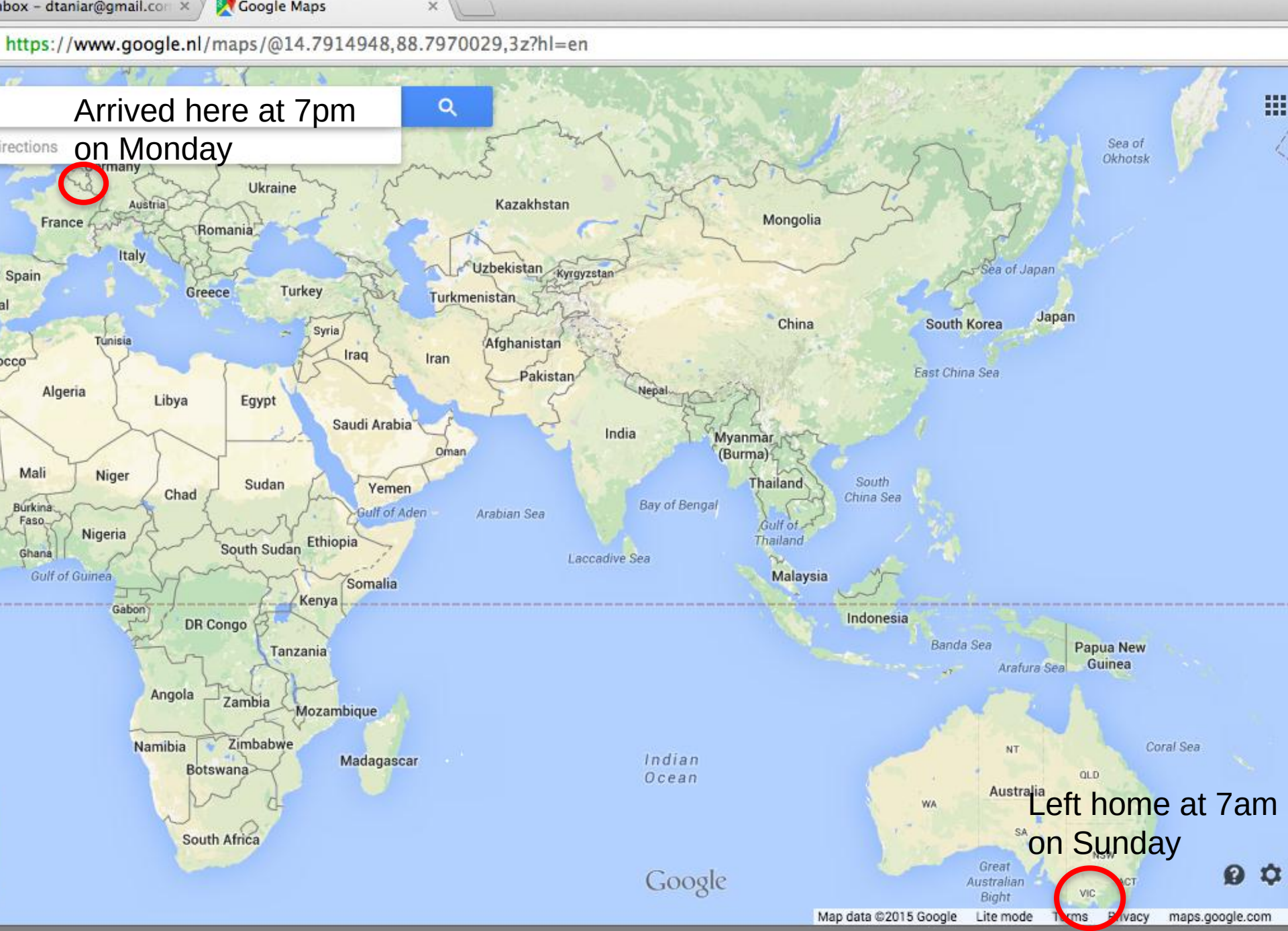


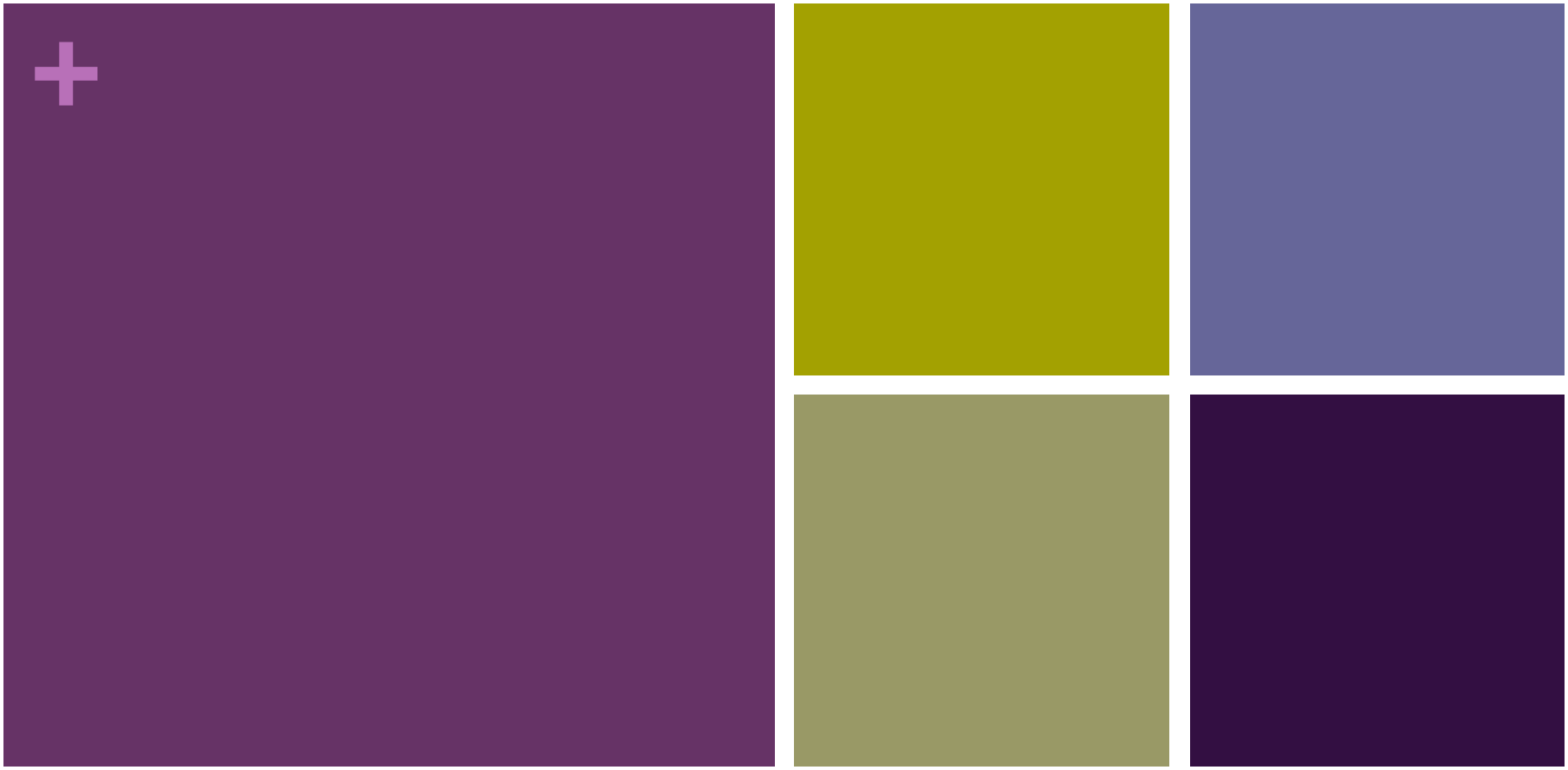
Transit time 9 hours

Transferred from Domestic to International


Jakarta to Amsterdam - Go
nl/maps/dir/Jakarta,+Special+Capital+Region+of+Jakarta,+Indonesia/Amsterdam/@28.3057542,43.8134911,3z/data=!4m14!4m13!1m5!1...
erlands 14 h 20 min
Transit time 6.5 hours
dtaniar@gmai
Amsterdam
Jakarta
14 h 20 min
Iceland
Norway
Finland
Russia
United Kingdom
Ireland
Denmark
Belarus
Ukraine
Austria
France
Romania
Italy
Spain
Portugal
Greece
Turkey
Kazakhstan
Mongolia
Uzbekistan
Kyrgyzstan
Turkmenistan
Sea of Japan
China
South Korea
Japan
East China Sea
Syria
Iraq
Tunisia
Morocco
Algeria
Libya
Egypt
Western Sahara
Saudi Arabia
Nepal
India
Myanmar (Burma)
Oman
Mauritania
Mali
Niger
Chad
Sudan
Yemen
Thailand
South China Sea
Bay of Bengal
Gulf of Aden
Arabian Sea
Gulf of Thailand
Burkina Faso
Guinea
Nigeria
Ghana
South Sudan
Ethiopia
Laccadive Sea
Malaysia
Gulf of Guinea
Somalia
Kenya
Gabon
DR Congo
Indonesia
Banda Sea
Tanzania
Arafura Sea
Angola
Zambia
Mozambique
Namibia
Zimbabwe
Botswana
Madagascar
Indian Ocean
Google
Map data ©2015 Google, INEGI
Lite mode
Terms
Privacy
maps.google.com
1000 km


dtaniar@gmail.com
Amsterdam to Luxembourg
https://www.google.nl/maps/dir/Amsterdam/Luxembourg+City,+Luxembourg/@51.0498988,3.8252892,7z/data=!4m14!4m13!1m5!1m1!1s0x47c63fb5949a77...
Netherlands—Luxembourg City, 1 h,from €183
Amsterdam
Schiphol Airport
Netherlands
Rotterdam
Belgium
Luxembourg City
Luxembourg
Paris
London
Brussels
Antwerp
Ghent
Cologne
Frankfurt
1 h
from €183
Google
Map data ©2015 GeoBasis-DE/BKG (©2009), Google
Lite mode
Terms
Privacy
maps.google.com
Report a problem
50 km


box - dtaniar@gmail.com
Google Maps
https://www.google.nl/maps/@14.7914948,88.7970029,3z?hl=en
Arrived here at 7pm on Monday
Left home at 7am on Sunday
Ukraine
Kazakhstan
Mongolia
Sea of Okhotsk
France
Austria
Romania
Italy
Spain
Greece
Turkey
Uzbekistan
Kyrgyzstan
Turkmenistan
Sea of Japan
China
South Korea
Japan
Syria
Iraq
Iran
Afghanistan
Pakistan
Nepal
East China Sea
Tunisia
Algeria
Libya
Egypt
Saudi Arabia
India
Myanmar (Burma)
Oman
Yemen
Thailand
South China Sea
Mali
Niger
Chad
Sudan
Burkina Faso
Nigeria
Ghana
Gulf of Aden
Arabian Sea
Bay of Bengal
Gulf of Thailand
Laccadive Sea
Gulf of Guinea
South Sudan
Ethiopia
Somalia
Kenya
Malaysia
Gabon
DR Congo
Tanzania
Indonesia
Banda Sea
Arafura Sea
Papua New Guinea
Angola
Zambia
Mozambique
Namibia
Zimbabwe
Botswana
Madagascar
Indian Ocean
Coral Sea
NT
QLD
Australia
WA
SA
NSW
ACT
VIC
Great Australian Bight
South Africa
Google
Map data ©2015 Google
Lite mode
Terms
Privacy
maps.google.com



Big Data is all about data that we don’t have

David Taniar
Monash University, Australia

EGC’2015, Luxembourg, 28 January 2015

1. Misconceptions
2. Motivating Examples
3. Challenges

Big Data is all about data that we don't have

David Taniar
Monash University, Australia

Big Data – a new buzzword

Call For Papers of conferences, workshops, journals in the period of 6 months (July-Dec last year)

Source: DBWorld

Big Data – a new buzzword

Call For Papers of conferences, workshops, journals in the period of 6 months (July-Dec last year)

Source: DBWorld

The 6th International Workshop on Workflow Management in Cloud and Big Data

3 - 5 Dec. 2014, Sydney, Australia

in conjunction with the International Conference on Big Data and Cloud Computing (BDCloud 2014)

Supported by IEEE TCSC Technical Area on Workflow Management in Scalable Computing Environments

CIKM 2014 Workshop on Interactive Mining for Big Data (ImBig)

Special Session on "Warehousing and Intelligent Analysis Of Complex Network Big Data (WIBIG 2014)" @DSAA'2014

1st International Workshop on Big Data Discovery & Curation

Co-located with KDD 2014

Sunday August 24, 2014, New York, USA

:::::::::::::::::::::::::::::: CALL FOR PAPERS ::::::::::::::::::::::::::::::

IEEE Transactions on Services Computing

Special Issue on Big Data Analytics, Infrastructure, and Applications

::

Pervasive and Mobile Computing

Special Issue on
Big Data Analytics for Smarter Health Care
- Call for Papers -

CFP: ACM TKDD Special Issue: Connected Health at Big Data Era

The availability of big data and the emergence of wearable computing, network science, and computational
landscape of how we decipher our lives, our social interactions, and our day-to-day activities. This well
transforming healthcare from reactive and hospital-centered, to preventive, proactive, evidence-based, pe
ailment recovery.

Call for Chapters: Managing Big Data in Cloud Computing Environments

2014 International Conference on Cloud Computing and Big Data

November 12-14, 2014 Wuhan, China

SIMBig'2014 - 1st Symposium on Information Management and Big Data
8-10 September 2014 - Cusco, PERU

Paper/Demos Submission Deadline: extended till July 11, 2014

http://www.lirmm.fr/simbig2014/
simbig2014@lirmm.fr

Big Data in Motion and Big Data at Rest (BD-MR)

In conjunction with the IEEE International Conference on Big Data, October 27, 2014, Washington DC, USA

3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data (BigSpatial-2014), Nov 4, 2014, Dallas, TX, USA.

Big Data in Motion and Big Data at Rest (BD-MR)

In conjunction with the IEEE International Conference on Big Data, October 27, 2014, Washington DC, USA

3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data (BigSpatial-2014), Nov 4, 2014, Dallas, TX, USA.

1st International Workshop on Privacy and Security of Big Data

(PSBD 2014)

in conjunction with

The 23rd ACM International Conference on Information and Knowledge Management (CIKM 2014)

November 7, 2014, Shanghai, China

The 3rd International Symposium on Privacy and Security in Cloud and Big Data (PriSec 2014)

3-5 December 2014, Sydney, Australia

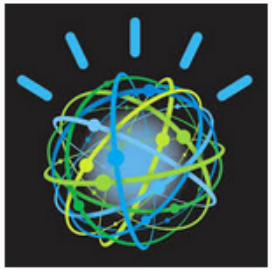
http://www.swinflow.org/confs/prisec2014

co-located with The 4th IEEE International Conference on Big Data and Cloud Computing (BdCloud2014)

Sponsored by IEEE TCSC Technical Area on Privacy and Security in Cloud and Big Data

IBM Research

Featured research | Cognitive computing | Client programs | Locations | Our people



Workshop on Enterprise Big Data Semantic and Analytics Modeling

Overview

IEEE BigData 2014

(CASK-14): Collaborative methodologies to Accelerate Scientific Knowledge discovery in big Data

International Workshop on Collaborative Big Data (C-Big 2014)

October 22, 2014 Miami, Florida, US

In conjunction with the IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2014)

Workshop on Semantics for Big Data on the Internet of Things (SemBIoT 2014)

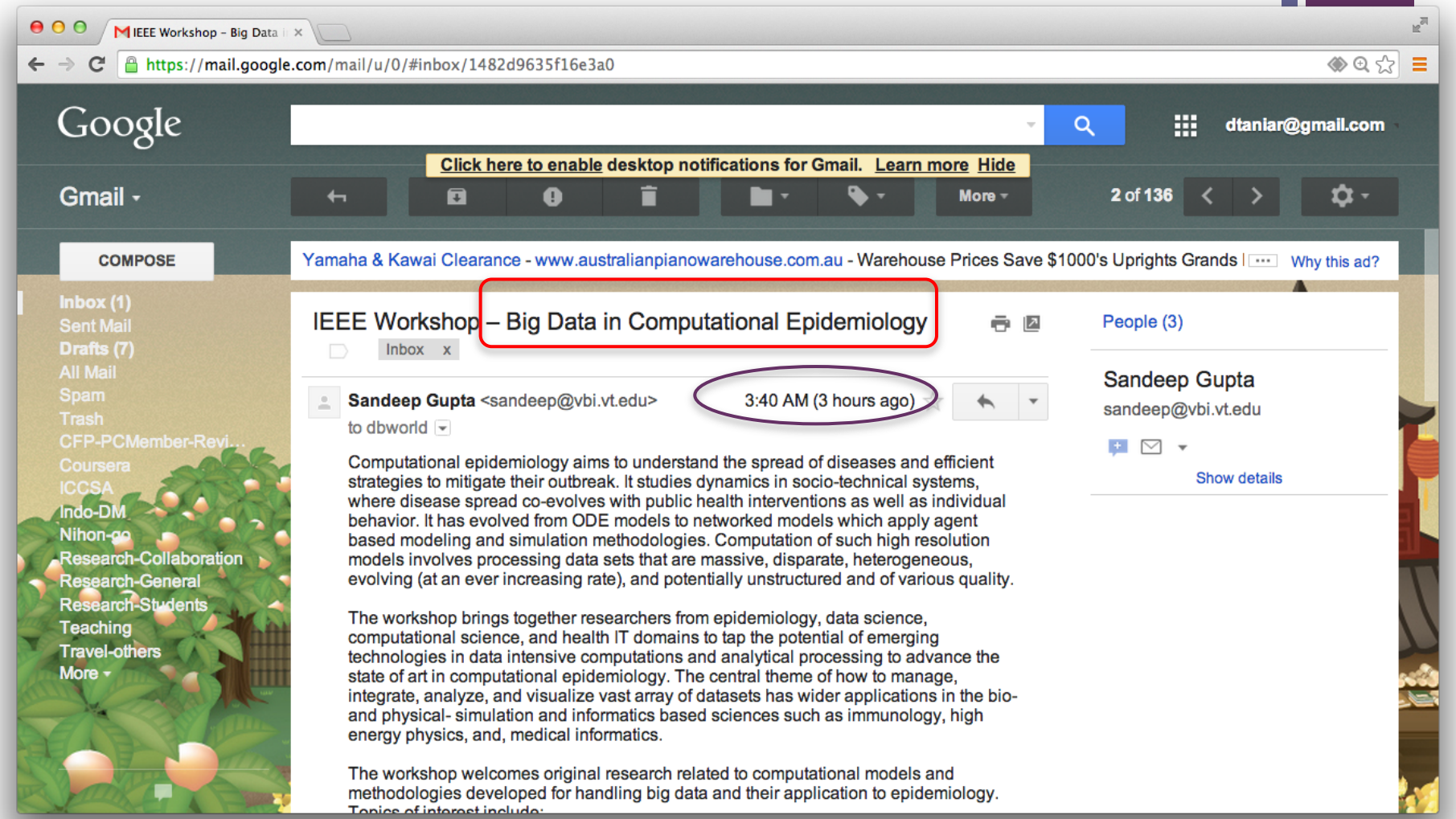
BigData in Bioinformatics and Healthcare Informatics

Washington D.C., Oct 27th, 2014 in conjuction with
The IEEE International Conference on BigData

BigData in Bioinformatics and Healthcare Informatics

Washington D.C., Oct 27th, 2014 in conjuction with
The IEEE International Conference on BigData

10110
01ACG
TTCGA



Big Data – a new buzzword

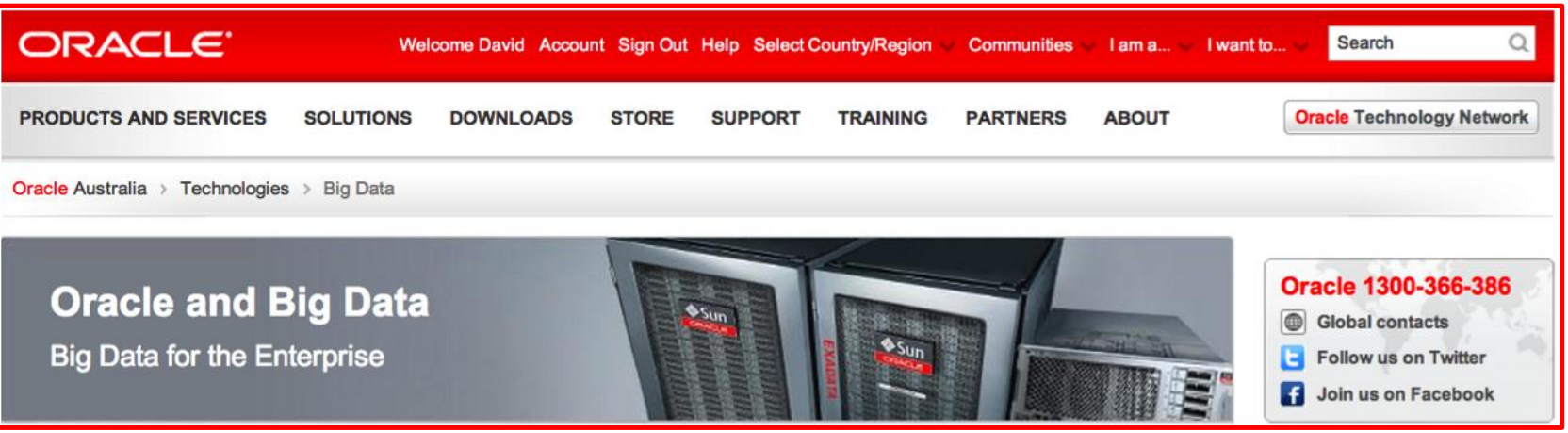
How many CFPs in Big Data are shown in the previous slides?

Big Data – a new buzzword

26

How many CFPs in Big Data are shown in the previous slides?

Big Data – a new buzzword



Big Data – a new buzzword

ORACLE

Welcome David Account Sign Out Help Select Country/Region Communities I am a... I want to... Search

PRODUCTS AND SERVICES SOLUTIONS DOWNLOADS STORE SUPPORT TRAINING PARTNERS ABOUT Oracle Technology Network

Oracle Australia › Technologies › Big D

Oracle and Big D

Big Data for the Enterpris

amazon web services

Sign Up My Account / Console English

AWS Products & Solutions AWS Product Information Developers Support

Big Data on AWS

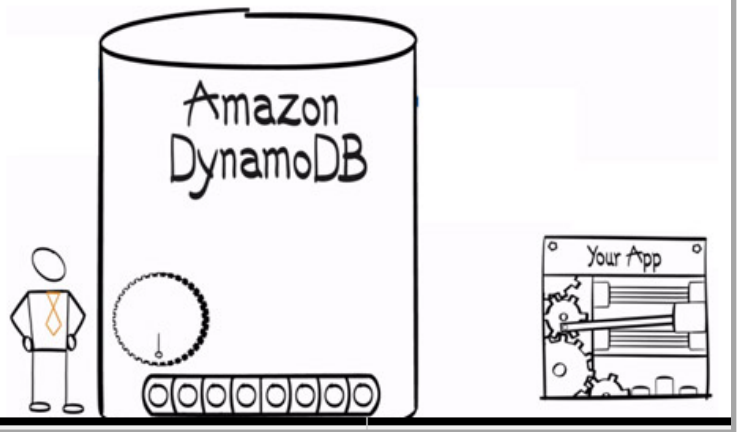
Drive innovation through data, with scalable services for data collection, storage, integration, analytics and collaboration.

Get started for free »

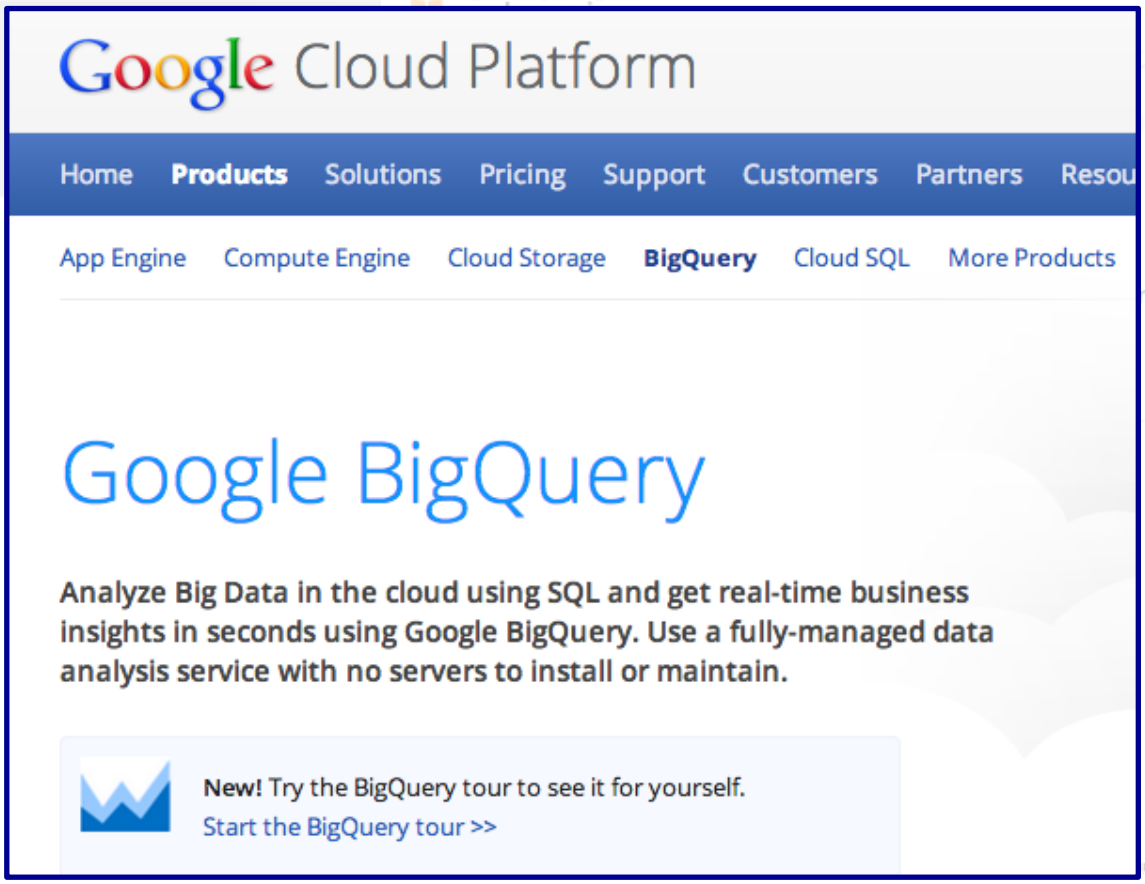
Learn about the AWS Free Tier

Amazon DynamoDB is a fully managed NoSQL database service that provides fast and predictable performance with seamless scalability.

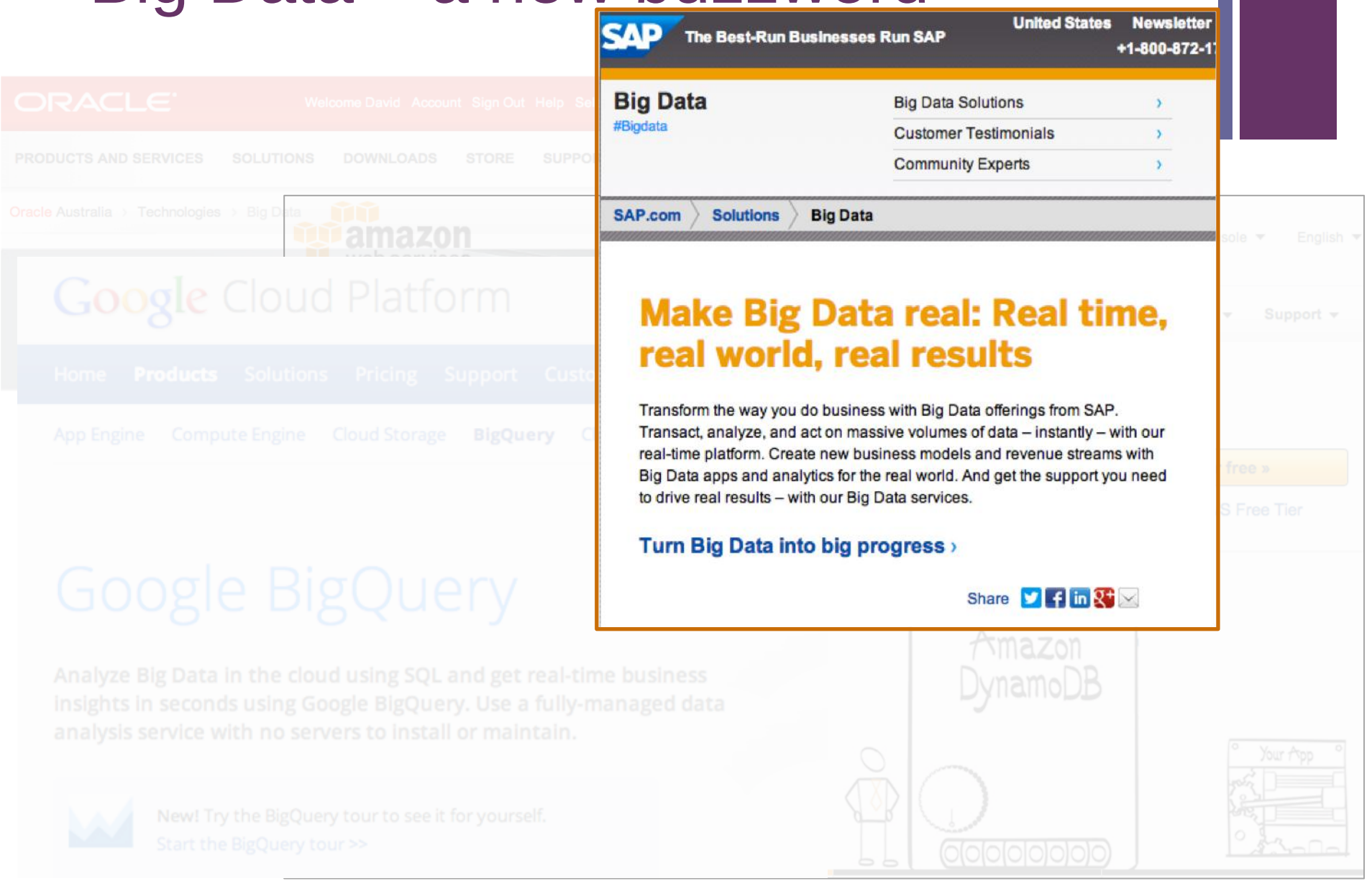
Learn more about DynamoDB »



Big Data – a new buzzword

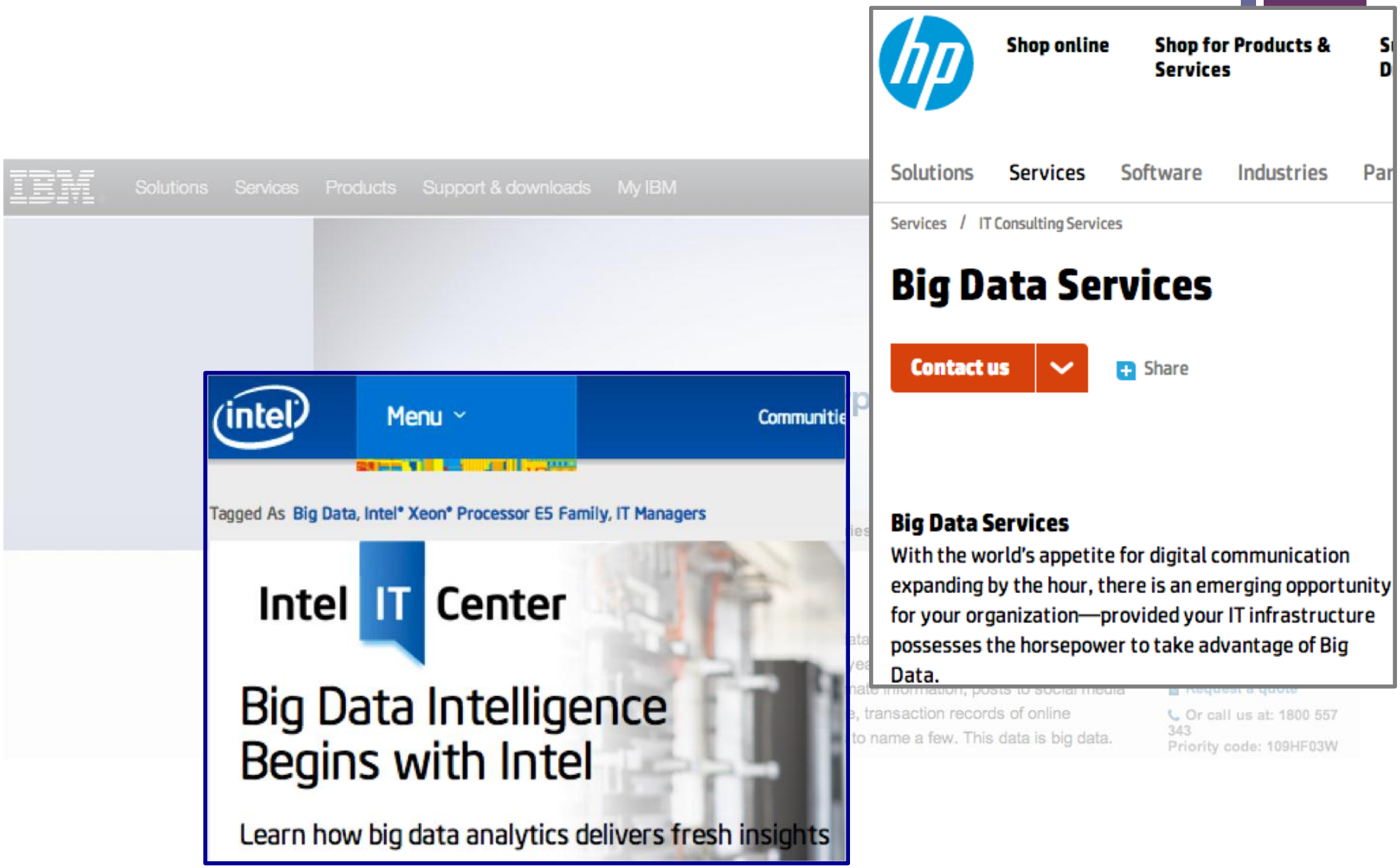


Big Data – a new buzzword

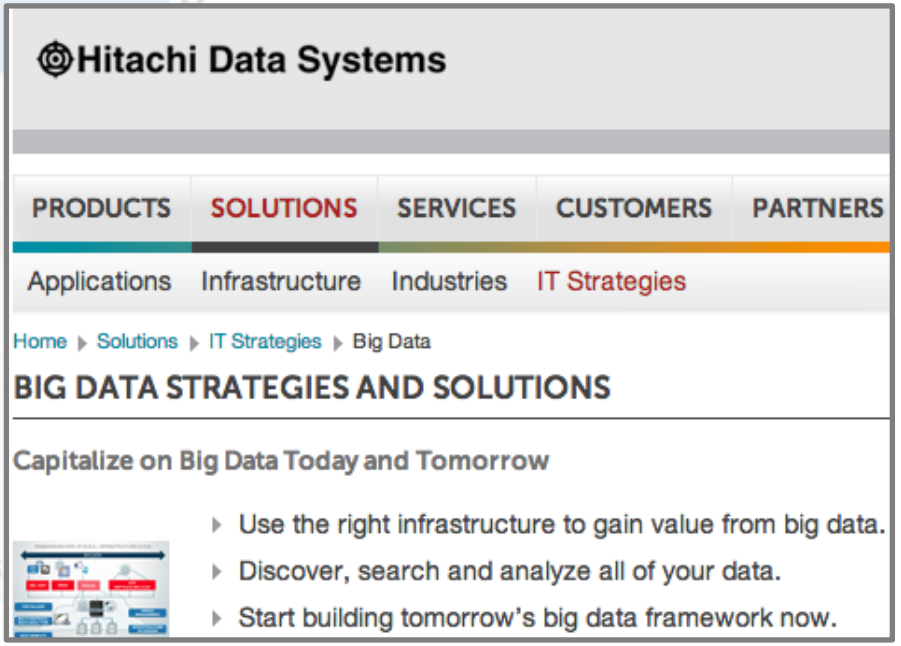


Big Data – a new buzzword

Big Data – a new buzzword



Big Data – a new buzzword



Big Data – a new buzzword

Big Data – a new buzzword

1. **Misconceptions**
2. **Motivating Examples**
3. **Challenges**

How big is “BIG”? – The Big Misconception

“**Everyday**, 2.5 **quintillion** bytes of data are created and 90% of the data in the world today was created within the past two years”.

IBM Corporation

10^6 = million (megabytes)
10^9 = billion (gigabytes)
10^{12} = trillion (terabytes)
10^{15} = quadrillion (petabytes)
10^{18} = **quintillion** (exabytes)

How big is “BIG”? – The Big Misconception

“**Everyday**, 2.5 **quintillion** bytes of data are created and 90% of the data in the world today was created within the past two years”.

IBM Corporation

“Worldwide information is more than **doubling every two years**, with **1.8 zettabytes** or 1.8 trillion gigabytes projected to be created and replicated this year alone”.

ZDNet news

…
10^{15} = quadrillion (petabytes)
10^{18} = **quintillion** (exabytes)
10^{21} = sextillion (**zettabytes**)

How big is “BIG”? – The Big Misconception

Data comes from everywhere:

- Post to **social media** sites

“As of September 2012, Facebook has more than **one billion** active users”

Facebook wikipedia

“Twitter has over 500 million users in 2012, generating over **340 million tweets** and handling over 1.6 billion search queries per day”.

Twitter wikipedia

How big is “BIG”? – The Big Misconception

Data comes from everywhere:

- Post to social media sites
- **Digital pictures** and **videos** posted online

“There has been more video uploaded to YouTube in the last 2 months than if ABC, NBC, and CBS had been airing content 24/7/365 continuously since 1948”.

Gartner Research

Facebook handles **50 billion photos** from its use base.

How big is “BIG”? – The Big Misconception

Data comes from everywhere:

- Post to social media sites
- Digital pictures and videos posted online
- **Transaction record** of online purchases

“Walmart handles more than **1 million customer transactions every hour**, which is imported into database estimated to contain more than 2.5 petabytes of data”.

The Economist

How big is “BIG”? – The Big Misconception

Data comes from everywhere:

- Post to social media sites
- Digital pictures and videos posted online
- **Transaction record** of online purchases

“Amazon's websites averaged **94 million unique visitors** per month, compared to 22 million visitors to Target’s website and 17 million visitors to Best Buy's online presence”.

Statista, Germany

How big is “BIG”? – The Big Misconception

Data comes from everywhere:

- Post to social media sites
- Digital pictures and videos posted online
- Transaction record of online purchases
- **Mobile phone** use and GPS signals

“There are **6 billion** mobile-phone subscriptions worldwide, and there are between 1 and 2 billion people accessing the Internet”.

The Economist

How big is “BIG”? – The Big Misconception

Data comes from everywhere:

- Post to social media sites
- Digital pictures and videos posted online
- Transaction record of online purchases
- Mobile phone use and GPS signals
- Various sensor and **satellite** data

“The Sloan Digital Sky Survey produced **13 petabytes** of data in one year”.

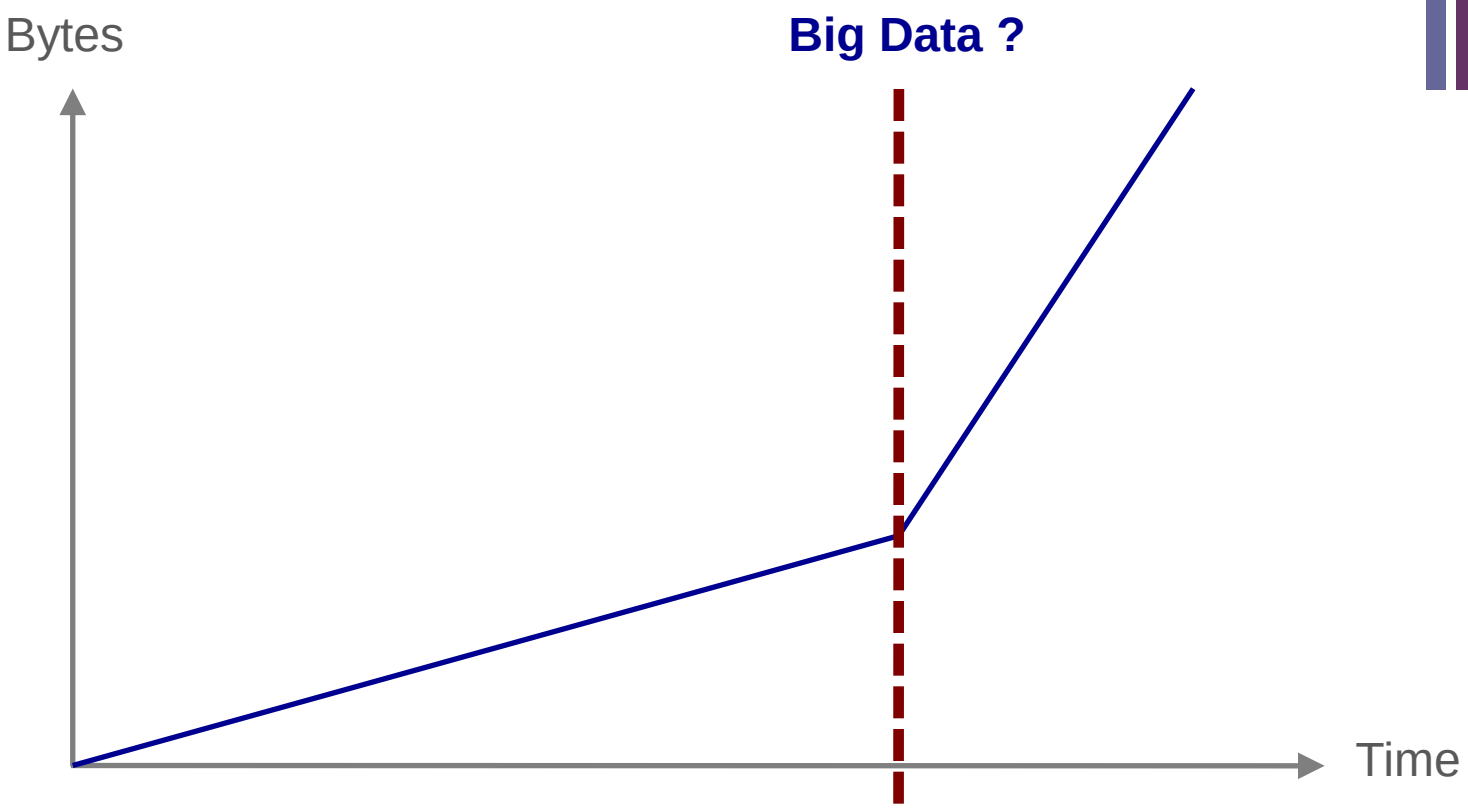
Nature

But… is it **Big Data**?

Are all these Big Data?

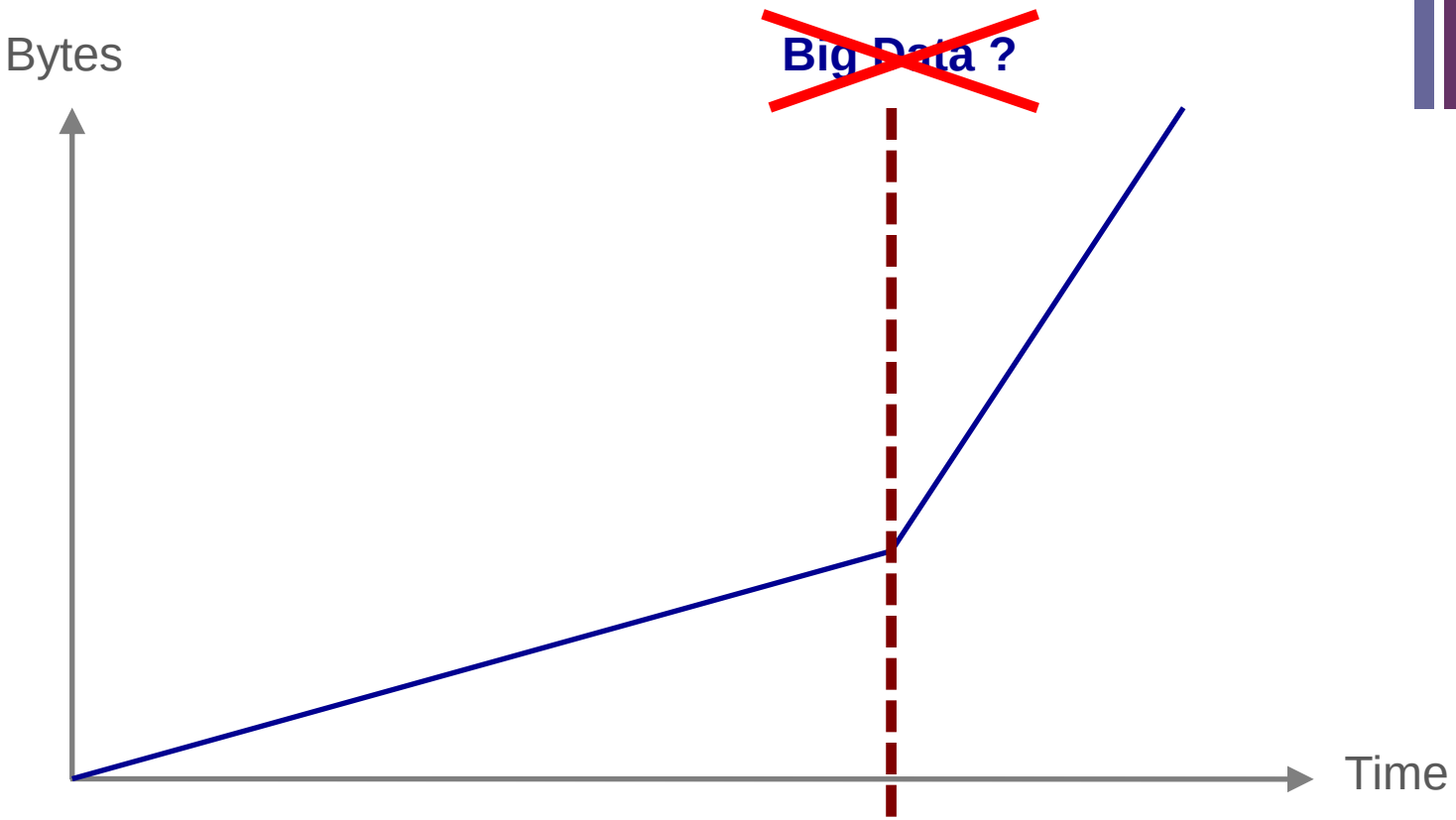
- Post to social media sites
- Digital pictures and videos posted online
- Transaction record of online purchases
- Mobile phone use and GPS signals
- Various sensor and satellite data

But… is it **Big Data**?



Misconception #1: Big Data is Big, but it is not about big data volume.

But… is it **Big Data**?



Misconception #1: Big Data is Big, but it is not about big data volume.

Is It Big Data Store?

Another Misconception

- If Big Data is not about Big Data Volume,

  Then… Is Big Data about data store?

- Big Data certainly needs big data storage.

  But… Is Big Data about data storage?

amazon web services

Sign Up | My Account / Console | English

AWS Products & Solutions | AWS Product Information | Developers | Support

Big Data on AWS

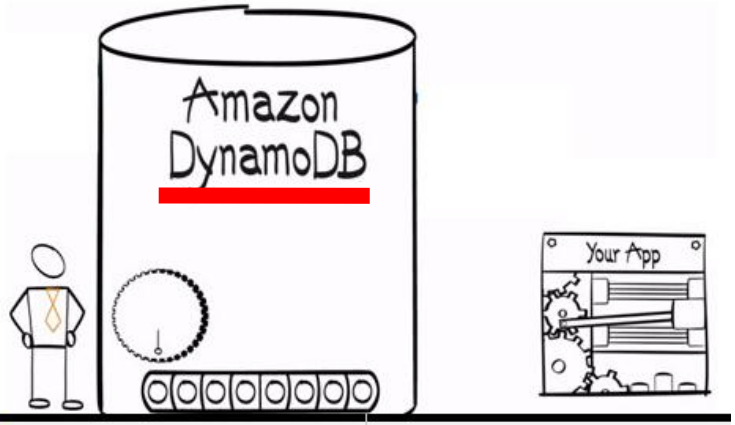
Drive innovation through data, with scalable services for data collection, storage, integration, analytics and collaboration.

Get started for free »

Learn about the AWS Free Tier

Amazon DynamoDB is a fully managed NoSQL database service that provides fast and predictable performance with seamless scalability.

Learn more about DynamoDB »



Hadoop at any scale

Zero-admin, NoSQL data storage

Quick start. High scale.

Name-your-price analytics

Capabilities

Big Data: end to end

"Big Data" refers to a collection of tools, techniques and technologies for working with data productively, at any scale. The tools to support data collection, computation along with collaboration and sharing are all available in a couple of clicks, with Amazon Web Services.

Learn more about Elastic MapReduce »

Is It Big Data Store?

Another Misconception

- Most data are in files:
  - Web is file centric
  - Email is file centric
  - Science is file centric
  - …
- **Big data is not in databases**
- DBMS delivers better performance than other systems
- So perhaps, a new database system, like **NoSQL**, is the answer

Is It Big Data Store?

Another Misconception

NOSQL, the new DBMS?

- NoSQL = **no-SQL** or **Not Only SQL.** Non-relational data storage systems.
- NoSQL movement was probably inspired by **Google's Big Table** or even **Amazon's S3**.
- Explosion of social media sites (Facebook, Twitter) with large data needs
- NoSQL feature: shared-nothing horizontal scaling – replicating and partitioning data over many servers.

Definition: "Next generation databases mostly addressing some of the points: being **non-relational**, **distributed**, **open-source**, and **horizontal scalable**. The original intention has been modern web-scale databases. The characteristics are: eventually consistent/BASE (not ACID), a huge data amount, and more."

www.nosql-database.org

Is It Big Data Store?

Another Misconception

- Varieties of NoSQL:
  - **Column-stores**: each storage block contains data from only one column
  - **Document stores**: stores documents made up of tagged elements
  - **Key-value stores**: hash table of keys

Facebook's Cassandra	MongoDB	Redis
Google's Big Table	CouchDB	Riak
Yahoo's PNUTS		Tokyo Cabinet
Apache's Hbase		Scalaris
Hypertable		Memcached, Membrain, Membase
HadoopDB		

Is It Big Data Store?

Another Misconception

- Leading users of NoSQL
  - Social networking sites (Twitter, Facebook, LinkedIn)
- Most corporate IT companies are not social networking industries.

Misconception #2: Big Data is not about Big Data Store, nor a new wave of DBMS, like NoSQL, although Big Data comes in a variety of data formats (e.g. social media formats)

- Although big companies have now been offering data storage for Big Data; and
- DBMS vendors are now selling data management systems for Big Data.

Is It Big Data Processing?

The Last Misconception

If Big Data were about big data volume, then we would need parallel processing to process Big Data.

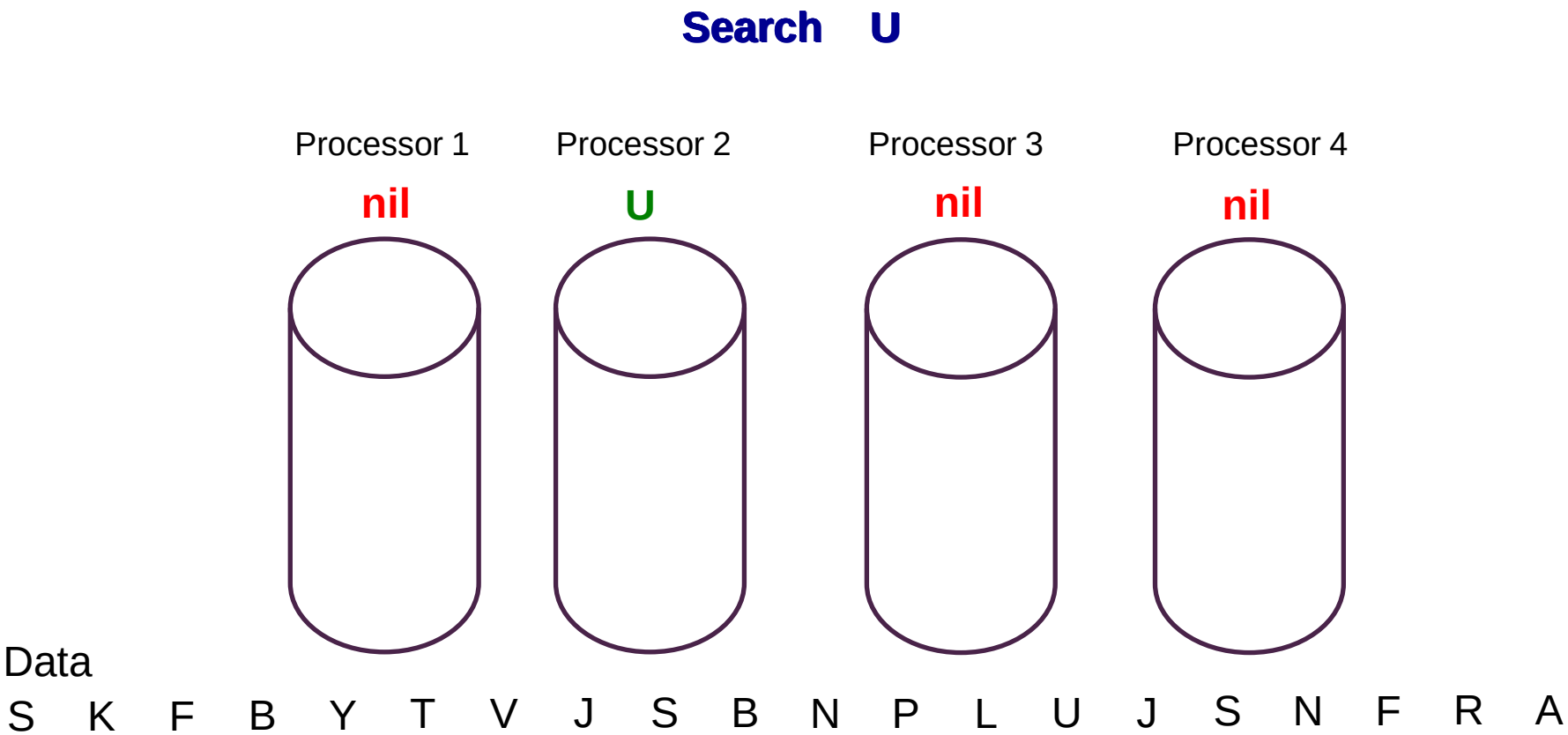
But…

- **Parallel Databases**: started in mid-late 80s.
- The trend was to build specialized database machines (e.g. Gamma, Bubba).
- Various parallel machine architectures were built:
  - Shared-memory architecture
  - Shared-nothing architecture
  - Shared-something architecture
- Due to the extensive research in parallel databases, now most commercial DBMSs (Oracle) have parallelization capabilities.

Is It Big Data Processing?

The Last Misconception

- Parallelization through data partitioning
- Hence, parallel scans, yield **I/O parallelism**



Is It Big Data Processing?

The Last Misconception

- Parallel Databases
- Then…
  - **Cluster Computing**
  - **Grid Computing**
  - **Cloud Computing**

Is It Big Data Processing?

The Last Misconception

- Parallel Databases
- Then…
  - Cluster Computing
  - Grid Computing
  - Cloud Computing
- And Now…
  - **MapReduce**
  - **Hadoop**

Big Data is often associated with MapReduce/Hadoop.
But, is Big Data about MapReduce/Hadoop?

amazon web services

Sign Up | My Account / Console | English

AWS Products & Solutions | AWS Product Information | Developers | Support

Big Data on AWS

Drive innovation through data, with scalable services for data collection, storage, integration, analytics and collaboration.

Get started for free »

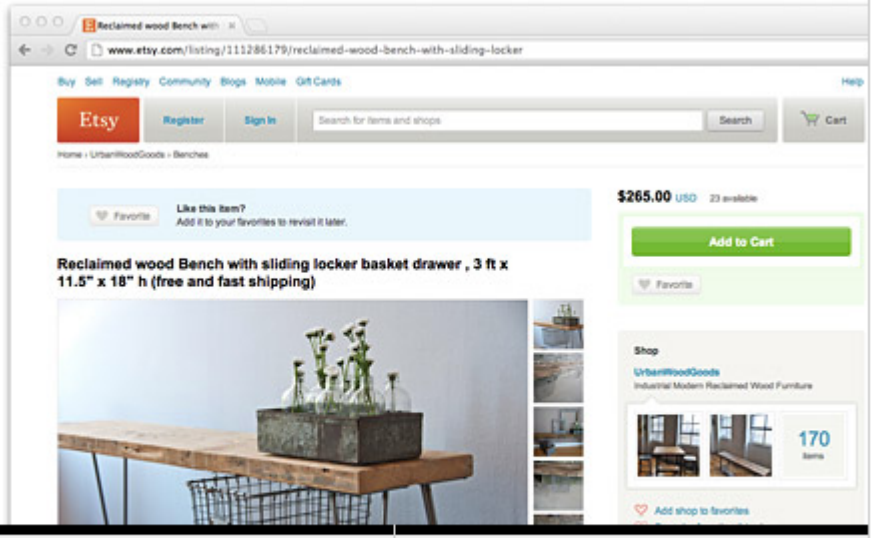
Learn about the AWS Free Tier

"Amazon Elastic MapReduce enables us to focus on our Hadoop-based analysis without worrying about the underlying infrastructure."

- Jason Davis, Director of Search & Personalization

Etsy

Read the Story »



Hadoop at any scale

Zero-admin, NoSQL data storage

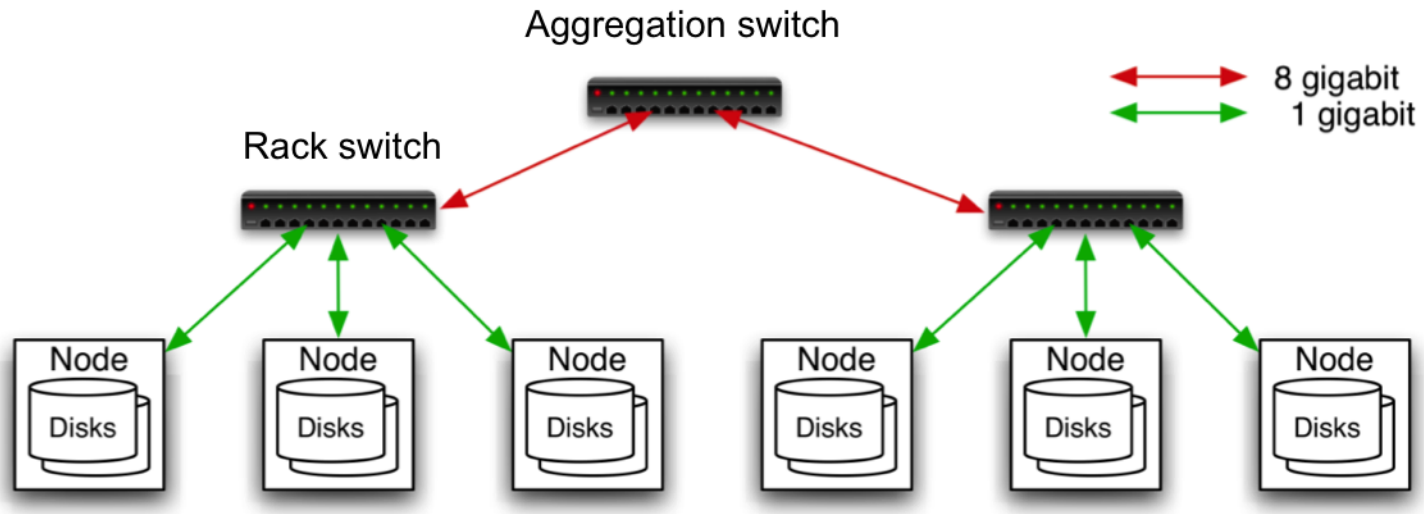
Quick start. High scale.

Name-your-price analytics

Is It Big Data Processing?

The Last Misconception

- Parallel computing
  - Constructing high performance parallel computers using a large number of (low-end) **commodity processors**.
  - Commodity machines (cheap, but unreliable).
  - Commodity network.
  - Scalable (1000's of machines, 10,000's of disks)



Is It Big Data Processing?

The Last Misconception

- Parallel programming
  - What happens with parallel programming that has existed for many decades (e.g. MPI)?
  - A new parallel programming paradigm: **MapReduce**
- MapReduce: a simple data-parallel programming model designed for **scalability** and **fault-tolerance**.
- Pioneered by Google
  - Processes 20 Petabytes of data per day
- Popularized by open-source Hadoop project
  - Used at Yahoo!, Facebook, Amazon, …

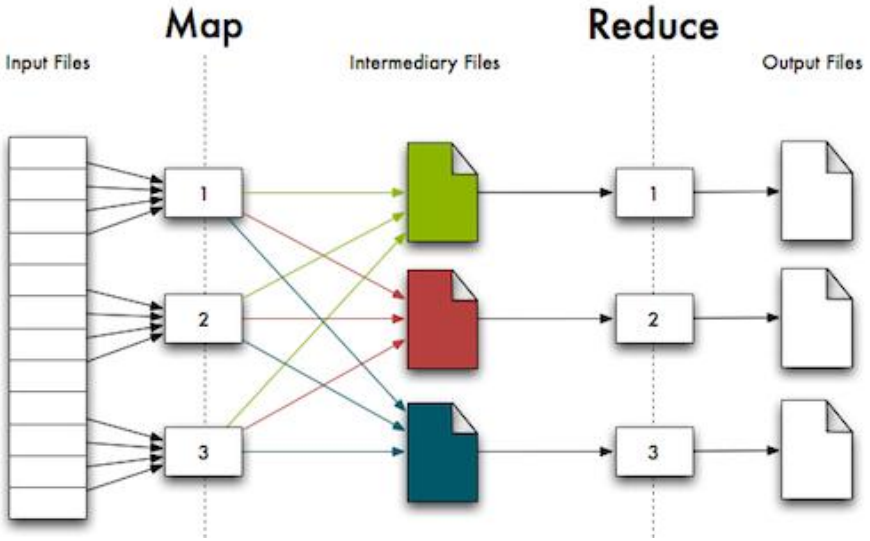
Is It Big Data Processing?

The Last Misconception

- **Cheap nodes fail**, especially if you have many of them
  - Mean time between failures for 1 node = 3 years
  - Mean time between failures for 1000 nodes = 1 day
  - ***Solution***: Build fault-tolerance into system
- **Commodity network = low bandwidth**
  - ***Solution***: Push computation to the data
- **Programming distributed systems is hard**
  - ***Solution***: Data-parallel programming model users write “map” and “reduce” functions, system distributes work and handles faults.

MapReduce

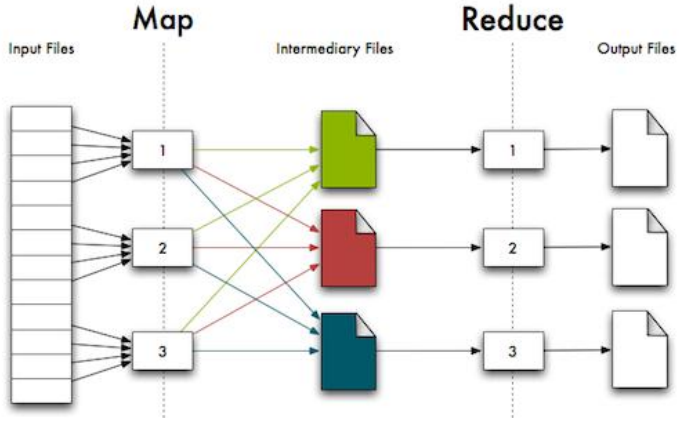
Simple Programming Model



- Single *master* controls job execution on multiple *slaves*
- Mappers preferentially **placed on same node** or same rack as their input block
  - Minimizes network usage
- Mappers save **outputs to local disk** before serving them to reducers
  - Allows recovery if a reducer crashes
- Allows having more reducers than nodesMappers save **outputs to local disk** before serving them to reducers
  - Allows recovery if a reducer crashes
  - Allows having more reducers than nodes

MapReduce

Fault Tolerance



1. If a **task crashes**, Retry on another node
   - OK for a map because it has no dependencies
   - OK for reduce because map outputs are on disk

2. If a **node crashes**:
   - Re-launch its current tasks on other nodes
   - Re-run any maps the node previously ran (Necessary because their output files were lost along with the crashed node)

3. If a **task is going slowly** (straggler):
   - Launch second copy of task on another node
   - Take the output of whichever copy finishes first, and kill the other

Surprisingly important in large clusters

- Stragglers occur frequently due to failing hardware, software bugs, misconfiguration, etc
- Single straggler may noticeably slow down a job

Is It Big Data Processing?

The Last Misconception

- Parallel Databases
- Then…
  - Cluster Computing
  - Grid Computing
  - Cloud Computing
- And Now…
  - **MapReduce**
  - **Hadoop**

Misconception #3: Parallel processing certainly helps. But Big Data processing is not about MapReduce/Hadoop or any parallel technologies.

The Big Data Misconceptions

1. It is not about Big Data Volume, although Big Data is big;
2. It is not about new Data Stores, although Big Data needs big data stores as data comes in a rich data formats; and
3. It is not about new trends in Parallel Processing, although parallel processing is needed to process Big Data.

So… what is then Big Data?

Big Data

1. **Misconceptions**
2. **Motivating Examples**
3. **Challenges**

Motivating Examples

1. A Clothing Store
2. An Airline
3. Health

Motivating Example – 1

- Sales of a **clothing shop** decreasing
- They analyzed their customers data and purchases to target certain customers
- Even after launching promotions, the sales was not improved

Motivating Example – 1

- Sales of a **clothing shop** decreasing
- They analyzed their customers data and purchases to target certain customers
- Even after launching promotions, the sales was not improved
- They also used **geofencing**
- And still unsuccessful
- **Puzzled** ???!!!!###%%%

Motivating Example – 1

- **Their own data did not and could not solve their sales problems.**
- So, they went out and searched for outside data – in this case social media tweets.
- They found from tweets from teenagers that the clothes currently being displayed for sales have the matched clothes from the previous seasons which are no longer available in the shop.

Motivating Example – 1

A Big Data Problem…

- Their own data did not and could not solve their sales problems.
- Because they did not have the necessary data to solve their sales problems.

Motivating Example – 2

An airline case study…

- Top level management of an airline company would like to know the service provided at the operational levels – to find out if their customers are happy or not
- Top level management is not able to see the operational levels. They only see the high level management.
- They can analyze their own data, and will not find anything.
- **A Big Data Problem** – their own data will not provide the solution to their problems.

Motivating Example – 2

An airline case study…

- So, they conducted a “sentiment analysis” from social media.
- To find out how the feeling from people toward the airline
- No questionnaires – so it is more objective
- Based on the results they get from sentiment analysis, they are able to rank the airlines based on the satisfactory from passengers through social media.
- Input (twitter feeds), output (ranking of airlines)

Motivating Example – 2

An airline case study…

- Sentiment analysis uses some quantitative values
- Long queue -1, nice food +1, delay -1, etc.
- Challenges:
  - It is supposed a negative comment, but the comment does not reflect the negativity explicitly
  - Use of punctuation (e.g. “good” ☹)

Motivating Example – 3

Big Data problem has been here long ago...

since 1800s !!!

Motivating Example – 3

Big Data problem has been here long ago...

since 1800s !!!

- There was **Cholera epidemic** in London (1831-1832).
- A lot of patients data has been collected. **Experts said** that cholera, like other diseases, was transmitted through inhalation of contaminated vapors.
- Cholera **struck England again** in 1854, even with more deaths; that proved the experts wrong.

Motivating Example – 3

Big Data problem has been here long ago...

since 1800s !!!

- There was **Cholera epidemic** in London (1831-1832).
- A lot of patients data has been collected. **Experts said** that cholera, like other diseases, was transmitted through inhalation of contaminated vapors.
- Cholera **struck England again** in 1854, even with more deaths; that proved the experts wrong.

Was it because they didn't have the technology? NO

Was it because they didn't have The data? YES

Motivating Example – 3

Big Data problem has been here long ago...

since 1800s !!!

A Big Data Problem!!!

- There was **Cholera epidemic** in London (1831-1832).
- A lot of patients data had been collected. **Experts said** that cholera, like other diseases was transmitted through inhalation of contaminated vapors.
- Cholera **struck England again** in 1854, even with more deaths.

Was it because they didn't have the technology? NO

Was it because they didn't have The data? YES

Motivating Example – 3

Big Data problem has been here long ago...

since 1800s !!!

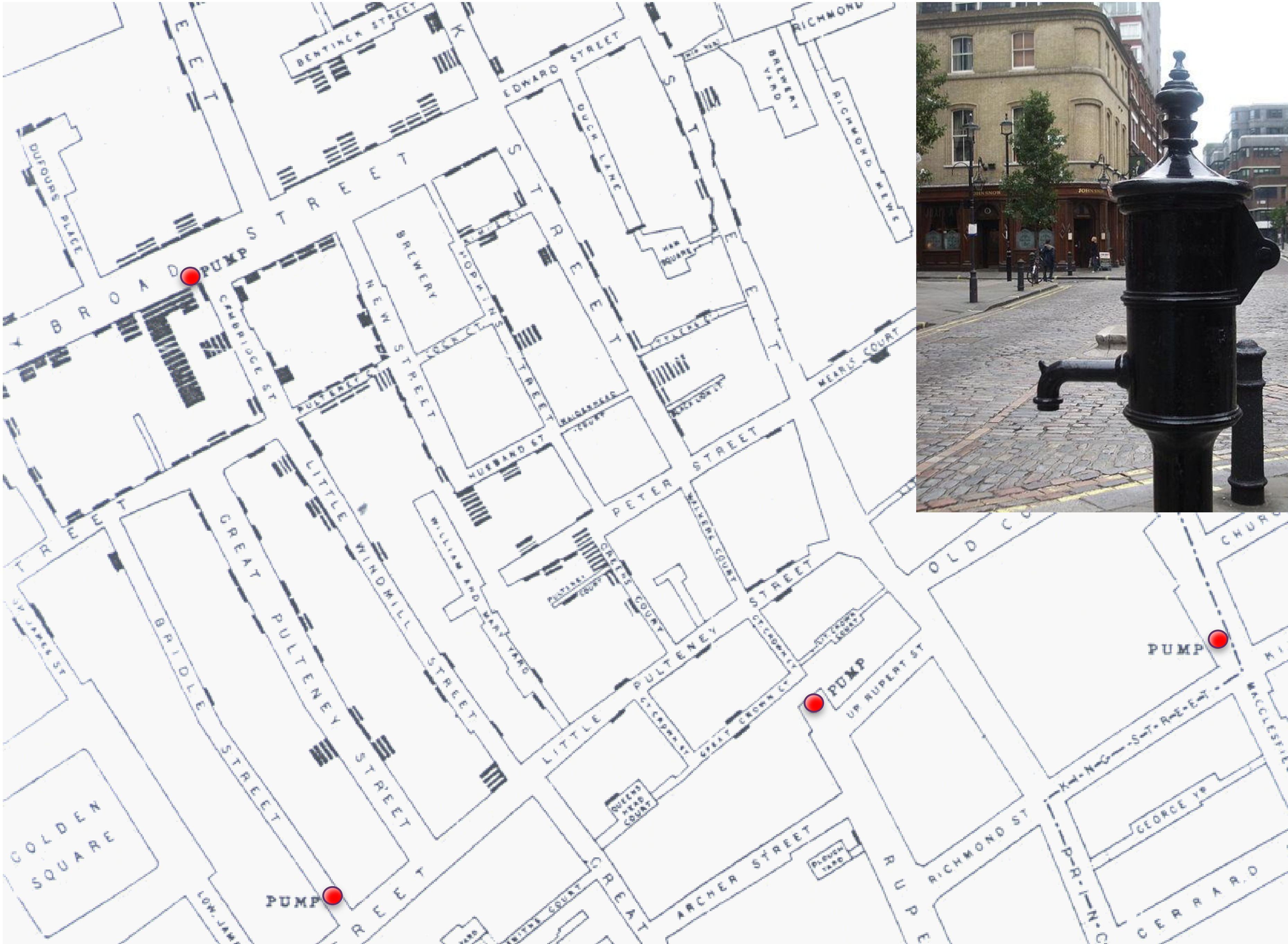
- **John Snow** was an apprenticed to a surgeon living at Newcastle-on-Tyne during the first cholera outbreak (1831-32). He then became a medical student, and later became a surgeon (MD, University of London, 1844).
- Dr Snow **published a paper** on “mode of communication of cholera”, suggesting that the “Cholera Poison” reproduced in the human body, and was spread through the contamination of food or water.
- Experts disagreed. Dr Snow could not prove them wrong, because the patient data they had did not indicate anything. They simply did not have the data. **Patient data was not enough!!!**

Motivating Example – 3

Big Data problem has been here long ago...

since 1800s !!!

- Then during the second Cholera outbreak in 1854, Dr Snow persisted, by collecting **non-health related data** of patients.
- He surveyed the location of the deaths and their **water pumps**.


BROAD STREET
PUMP
DUFOURS PLACE
BENTINCK STREET
EDWARD STREET
RICHMOND
BREWERY YARD
RICHMOND MEWS
DUCK LANE
CAMBRIDGE ST
NEW STREET
BREWERY
HOPKINS STREET
HUSBAND ST
PETER STREET
WALKERS COURT
MEARS COURT
GREAT PULTENEY STREET
LITTLE WINDMILL STREET
WILLIAM AND MARY YARD
LITTLE PULTENEY STREET
BRIDLE STREET
GOLDEN SQUARE
PUMP
PUMP
UP RUPERT ST
GREAT CROWN CT
QUEENS HEAD COURT
ARCHER STREET
RICHMOND ST
GEORGE YD
KING STREET
MACCLESFIELD
PUMP

Motivating Example – 3

http://www.youtube.com/watch?v=TUTmg1iVX8E

Big data is all about data that we don’t have…

Big Data

1. **Misconceptions**
2. **Motivating Examples**
3. **Challenges**

Big Challenges

1. Understanding Business Needs

- One very important question:

  Where is the data???

Also…

What data I need?

Who owns the data?

How to get the data?

Big Challenges

1. Understanding Business Needs

- Data is coming at a rate faster than what we can absorb.
- Data is coming from everywhere:
  - Social networks
  - Mobile phones
  - Financial data
  - Sensors
  - …
- Ambient data: data is everywhere, and yet we didn't see it.
- BUT… those who can react quickly wins.
- The importance is in the speed of the feedback loop, taking data from input to decision.

Big Challenges

2. Data Integration

Data Integration, or Data Linkage

- M Brodie (AINA2010 keynote): “Data integration accounts for ~**40%** software project costs”.
- Data Linkage for:
  - Health records
  - Spatial data linkage
  - …

Big Challenges

2. Data Integration

Data Cleaning and Data Preparation

- Raw data, not ready for any kind of processing and analytics.
- Not many work in this area.
- Perhaps not scientific enough, or considered dirty work,…
- …although data cleaning and data preparation account for **80%** of the entire data processing and analytics. It is a serious business.
- There are actually many available techniques: Parsing, data transformation, duplicate elimination, statistical methods, etc.

Big Challenges

3. Data Storage and Management

1. Data from social media, data streams…

Ella @MrsEllaBella Mar 7
Last night I discovered that @Cinnabon can be delivered throughout the UK. I dropped so many hints to hubby. Bet he didn't get it at all.
Expand

Lady Hannah @mardycow Mar 7
@MrsEllaBella @cinnabon @garethemery has promised to get me one of these when I visit him in LA!
Expand

Ella @MrsEllaBella Mar 7
@mardycow @**cinnabon** @garethemery they are amazing. I always get one at Chicago **airport** when transferring to/from Austin - love them!!
Hide conversation Reply Retweet Favorite More
12:34 p.m. - Mar 7, 2013 · Details

Cinnabon @Cinnabon Mar 7
@MrsEllaBella @mardycow We look forward to seeing you both! Until then, something to remember us by :) pic.twitter.com/DXysPJjzF9
View photo

Big Challenges

3. Data Storage and Management

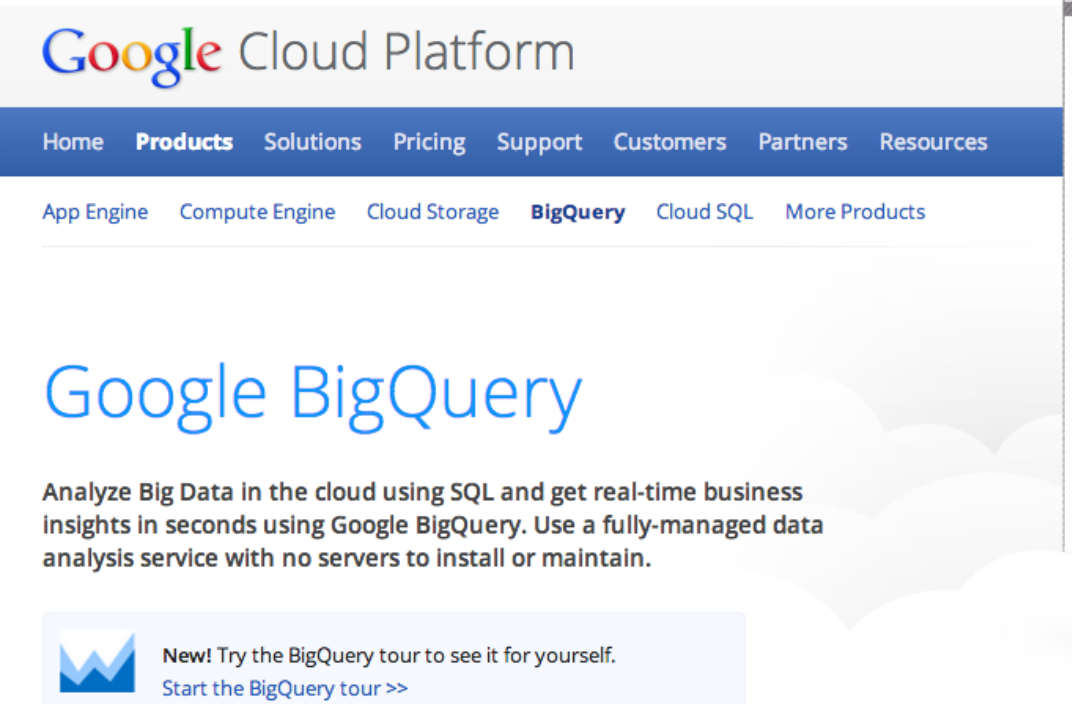
1. Data from social media, data streams…
2. DBMS, cloud storage, data storage technologies

Memory is the new disk

- Cost of main memory is plunging, so main memory solution becomes very practical.
- If the data not quite fits into main memory, you might look at flash as another technology.
- SAP's Main Memory Database – push application logic close to the database to leverage the sophisticated optimization and parallelization techniques.

Big Challenges

4. Analytics



Big Challenges

4. Analytics

- Is data integration a separate process?
- Will the integrated data remain decentralized, or become centralized?
- Will it be real-time analysis?
- Is the imported data preserve privacy?
- Will the processing unit capable to handle big data volume?

Big Data, Big Challenges

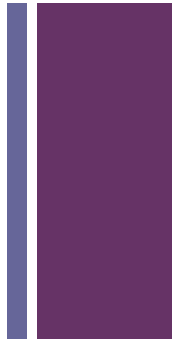
Conclusions…

1. Understanding Business Needs
2. Data Integration
3. Data Storage and Management
4. Analytics

Big Data is all about data that we don't have.

Other than this, we just do our research normally – and it is not Big Data.

Thank you…



Big Data is all about data that we don't have.

Other than this, we just do our research normally – and it is not Big Data.