Anthony Ventresque

anthony.ventresque@ucd.ie

Room B2.14 (CS building)

Big Data Programming

COMP47470

Introduction to COMP47470

School of Computer Science, UCD

Scoil na Ríomheolaíochta, UCD

Who Am I?

- Anthony Ventresque
  - B2.14 (CS Building)
- BSc and MSc in Philosophy, BSc, MSc and PhD in Computer Science (France), Research positions in NTU (Singapore), UCD and IBM Ireland

- Research Areas:
  - Parallel Data Management (=Big Data)
  - Software Quality (Testing)
  - Data Centre Optimisation (e.g., the Cloud)
  - Distributed Simulation of urban traffic

IBM skype Microsoft lero THE IRISH SOFTWARE RESEARCH CENTRE

Contact Details

anthony.ventresque@ucd.ie

- The best way to contact me is by email.
- Feel free to get in touch with questions.
- Don’t expect immediate responses, I will endeavour to respond within 48 hours.
- When emailing please state:
  - Your name (as it appears on Moodle),
  - Your class (COMP47470), and
  - Your student number.
- Use the forum, if you have a question you think the class would like to know the answer.

Outline

- Structure of the Module
- What is Big Data?

Take home message:

This module will give you the necessary understanding and some tools to store, ingest and process large, fast and complex data.

Module Aims

- What we want to give you in this module is:
  - the understanding of ***what is Big Data*** and why it is different from classical data management models
  - the ability to run some of the most important tools and platforms of the ***Hadoop galaxy***
  - the understanding of ***data collection and storage***
  - the understanding of how to ***process data efficiently***
  - the ability to select the ***right solution*** for your problem/your data

Syllabus

- Managing data at scale
- Understand the various data management paradigms (SQL/NoSQL)
- Manage your own cluster (Hadoop/HDFS)
- Understand big data programming models (MapReduce, Spark)
- Use Big Data solutions for streams of graphs
- Use Big Data solutions for streams of data

Lectures and Assessment

- 2 lectures per week
- 1 quiz each week (***10%***)
- 1 lab session per week
  - TA: Leandro Batista de Almeida & Ersi Ni
  - 3 projects/assignments (***50%*** total)
- 1 final exam (***40%***)

Grading

https://www.cs.ucd.ie/Grading/

Grade	Min	Max	Average
A+	95	100	97.5
A	90	95	92.5
A-	85	90	87.5
B+	80	85	82.5
B	75	80	77.5
B-	70	75	72.5
C+	65	70	67.5
C	60	65	62.5
C-	55	60	57.5
D+	50	55	52.5
D	45	50	47.5
D-	40	45	42.5
E+	35	40	37.5
E	30	35	32.5
E-	25	30	27.5
F+	20	25	22.5
F	15	20	17.5
F-	10	15	12.5
G+	8	10	9
G	5	8	6.5
G-	2	5	3.5
NG	0	0	0

Plagiarism & UCD Computer Science

- Plagiarism is a serious academic offence
  - [Student Code, section 6.2] or [UCD Registry Plagiarism Policy] or [CS Plagiarism policy and procedures]
- Our staff and demonstrators are proactive in looking for possible plagiarism in all submitted work
- Suspected plagiarism is reported to the CS Plagiarism subcommittee for investigation
  - Usually includes an interview with student(s) involved
  - 1st offence: usually 0 or NG in the affected components
  - 2nd offence: referred to the University disciplinary committee
- Student who enables plagiarism is equally responsible
  - http://www.ucd.ie/registry/academicsecretariat/docs/plagiarism_po.pdf
  - http://www.ucd.ie/registry/academicsecretariat/docs/student_code.pdf
  - http://libguides.ucd.ie/academicintegrity

Brightspace

- All lectures notes and lab materials will be available on Brightspace.
- Solutions will be provided to assessments, weekly quizzes, practicals BUT only after they happen.

- Module “name”: COMP47470-Big Data Programming 2018/19 Semester 2

Workload

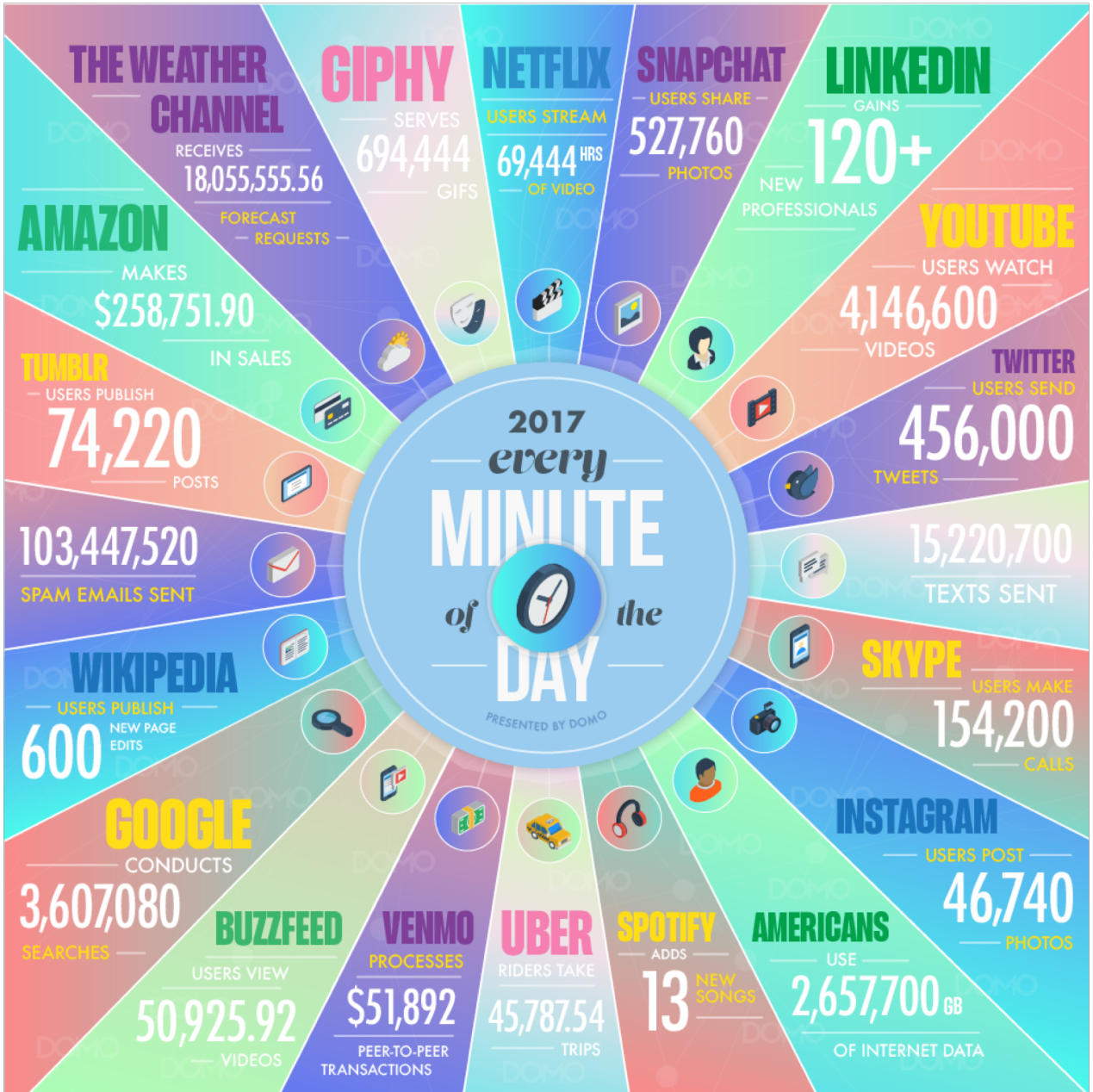
- Lectures
  - ~1 hour/lecture
  - Weekly Quizzes (5-15 minutes)
- Lab sessions
  - guided sessions, self contained
  - Projects/assignments: substantial amount of work required (~10 hours each)

What is Big Data?

- The three Vs:
  - Volume
  - Velocity
  - Variety
  - (Veracity)
  - …

What is Big Data?

https://www.domo.com/learn/data-never-sleeps-5



What is Big Data?

- 44 billion SQLs
- 23,000+ transactions/sec
- 2,000,000,000 transactions/day
- 5 billion API calls
- 150+ millions of active users

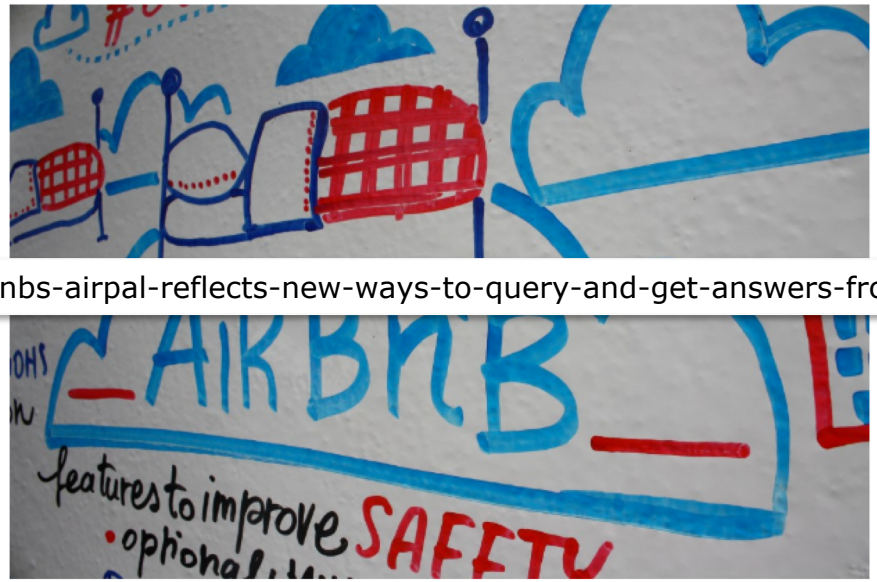
What is Big Data?

THENEWSTACK Technology Culture Tutorials Ebooks Podcasts Events

TECHNOLOGY / GLOBAL

Airbnb's AirPal Reflects New Ways to Query and Get Answers from Hive and Hadoop

9 Mar 2015 4:06pm, by Scott M. Fulton III

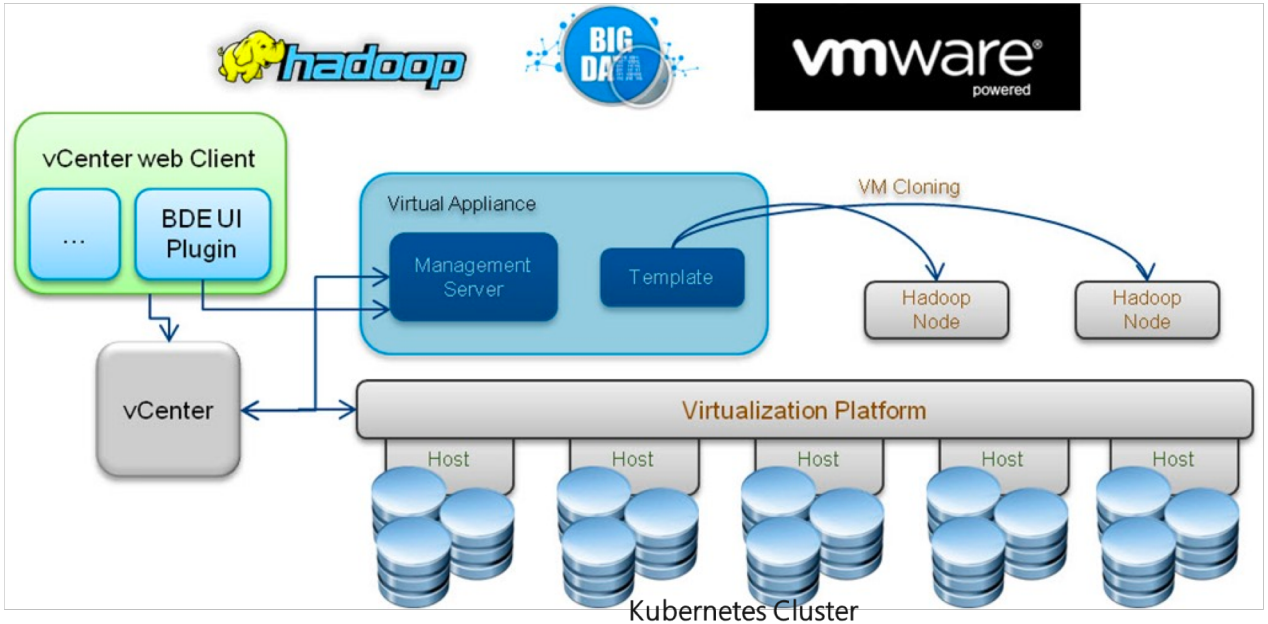


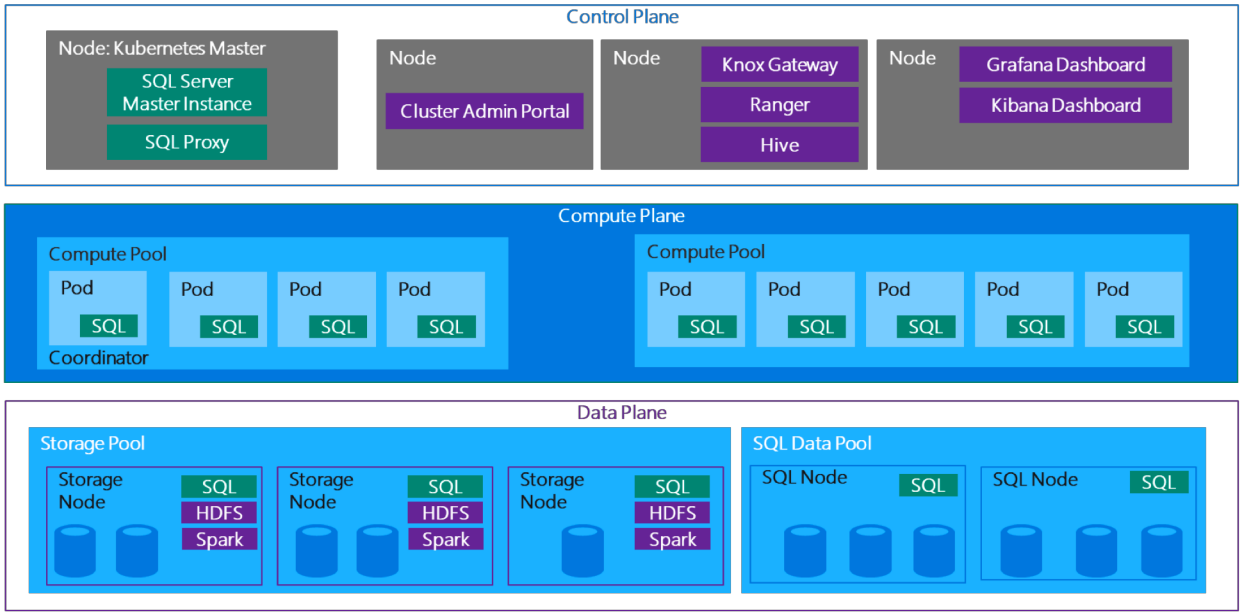
http://thenewstack.io/airbnbs-airpal-reflects-new-ways-to-query-and-get-answers-from-hive-and-hadoop/

Airbnb's data stores are approaching 1.5 petabytes in accumulated size — a mere drop in the bucket compared to Facebook's 300 petabytes, but a colossus, nonetheless. When Airbnb needed a tool for querying and visualizing that less-than-infinite pool of data, it built one for itself called AirPal. The visualization tool, which it has open sourced, examines records from clusters numbering in the tens of

What is Big Data?

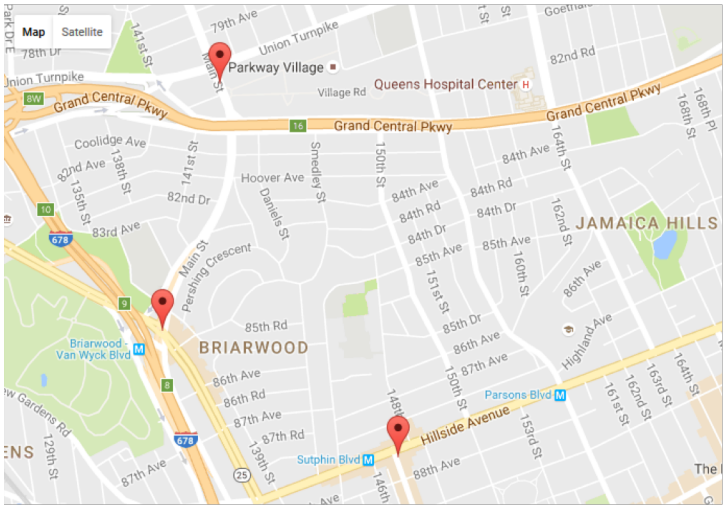
What is Big Data?



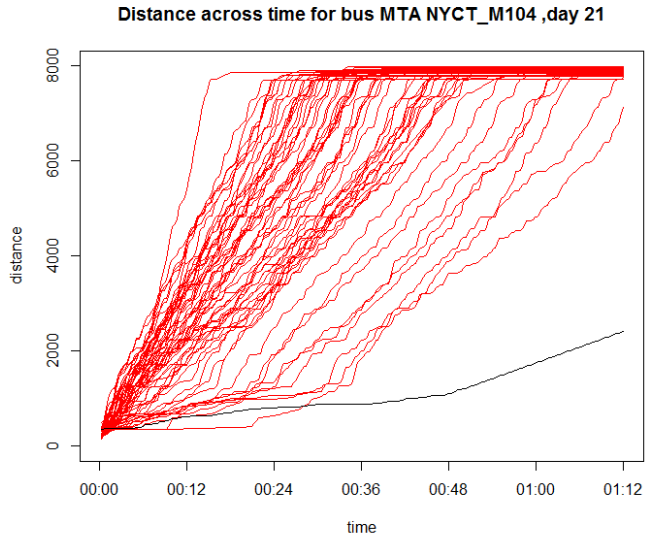


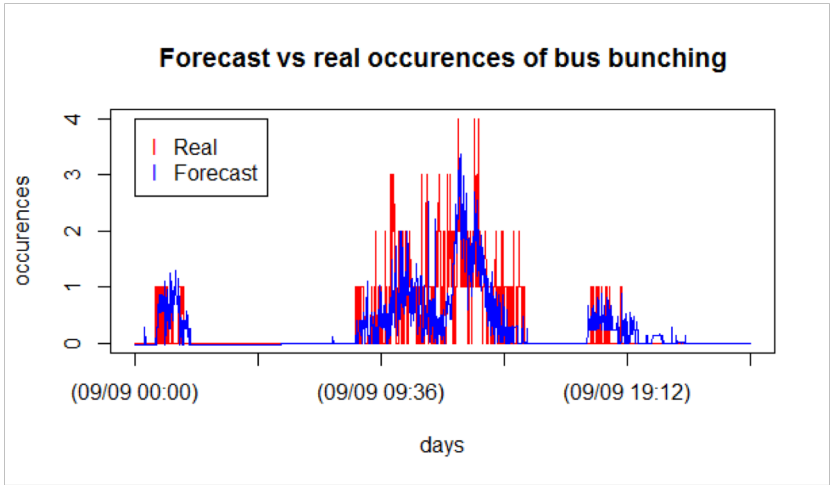
UCD DUBLIN

What is Big Data?

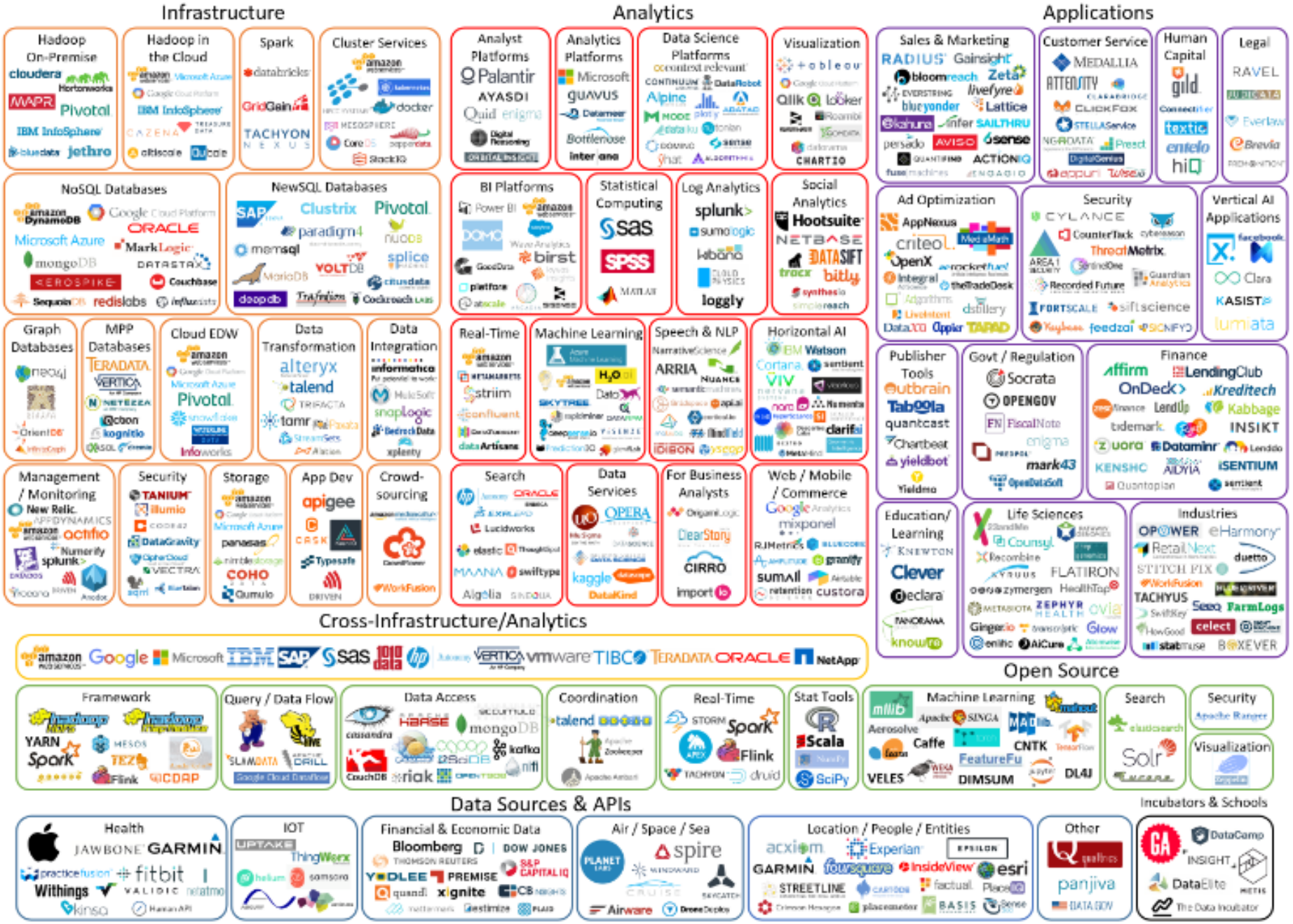


	latitude	longitude	time_received	vehicle_id	distance_along_trip	inferred_direction_id	inferred_phase	inferred_route_id
1	40.80980	-73.96243	(14-09-01 04:00:02)	5201	6628.672463	0	IN_PROGRESS	MTA NYCT_M104
2	40.81489	-73.95455	(14-09-01 04:00:04)	3824	7820.595206	0	LAYOVER_DURING	MTA NYCT_M104
3	40.75663	-73.99030	(14-09-01 04:00:12)	5205	4.339865	0	IN_PROGRESS	MTA NYCT_M104
4	40.75622	-73.98958	(14-09-01 04:00:14)	3804	8777.551234	1	LAYOVER_DURING	MTA NYCT_M104
5	40.76561	-73.97980	(14-09-01 04:00:15)	3925	7259.517107	1	IN_PROGRESS	MTA NYCT_M104
6	40.79135	-73.97428	(14-09-01 04:00:29)	3816	4004.392259	1	IN_PROGRESS	MTA NYCT_M104
7	40.80648	-73.96503	(14-09-01 04:00:31)	6743	2133.425217	1	IN_PROGRESS	MTA NYCT_M104
8	40.81145	-73.96120	(14-09-01 04:00:34)	5201	6836.881484	0	IN_PROGRESS	MTA NYCT_M104
9	40.81489	-73.95455	(14-09-01 04:00:37)	3824	7820.595206	0	LAYOVER_DURING	MTA NYCT_M104
10	40.75710	-73.98997	(14-09-01 04:00:44)	5205	64.607973	0	IN_PROGRESS	MTA NYCT_M104
11	40.75622	-73.98958	(14-09-01 04:00:46)	3804	8777.551234	1	LAYOVER_DURING	MTA NYCT_M104
12	40.76365	-73.98111	(14-09-01 04:00:47)	3925	7511.532536	1	IN_PROGRESS	MTA NYCT_M104
13	40.78907	-73.97594	(14-09-01 04:01:02)	3816	4293.243011	1	IN_PROGRESS	MTA NYCT_M104
14	40.80470	-73.96627	(14-09-01 04:01:03)	6743	2354.034063	1	IN_PROGRESS	MTA NYCT_M104
15	40.81464	-73.95886	(14-09-01 04:01:05)	5201	7244.464823	0	IN_PROGRESS	MTA NYCT_M104
16	40.81489	-73.95455	(14-09-01 04:01:11)	3824	7820.595206	0	LAYOVER_DURING	MTA NYCT_M104
17	40.75877	-73.98877	(14-09-01 04:01:15)	5205	271.544534	0	IN_PROGRESS	MTA NYCT_M104
18	40.75622	-73.98958	(14-09-01 04:01:17)	3804	8777.551234	1	LAYOVER_DURING	MTA NYCT_M104
19	40.76276	-73.98187	(14-09-01 04:01:18)	3925	7630.781087	1	IN_PROGRESS	MTA NYCT_M104
20	40.80435	-73.96652	(14-09-01 04:01:34)	6743	2402.136128	1	IN_PROGRESS	MTA NYCT_M104
21	40.78901	-73.97592	(14-09-01 04:01:36)	3816	4305.699499	1	IN_PROGRESS	MTA NYCT_M104
22	40.81491	-73.95868	(14-09-01 04:01:36)	5201	7276.643027	0	IN_PROGRESS	MTA NYCT_M104
23	40.81489	-73.95455	(14-09-01 04:01:44)	3824	7820.595206	0	LAYOVER_DURING	MTA NYCT_M104
24	40.75961	-73.98816	(14-09-01 04:01:46)	5205	383.933114	0	IN_PROGRESS	MTA NYCT_M104





Big Data Landscape 2016 (Version 2.0)



Last Updated 2/12/2016

FIRSTMARK

What is Big Data?

- The three Vs:
  - Volume
  - Velocity
  - Variety
  - (Veracity)

  - ***Processing and managing Big Data is skill- and labour-intensive***

Conclusion

- Volume
- Velocity
- Variety
- (Veracity)
- Processing and managing Big Data is skill- and labour-intensive

- Big Data Programming will give you the tools to store, ingest and process large, fast and complex data