Databases & Data Mining

Erwin M. Bakker & Stefan Manegold

e.m.bakker s.manegold
@liacs.leidenuniv.nl

https://homepages.cwi.nl/~manegold/DBDM/
http://liacs.leidenuniv.nl/~bakkerem2/dbdm/

Universiteit Leiden

DBDM: Overview

Period: September 11th - December 4th 2018 (Tuesdays)
Place: Room **312** (LIACS, Snellius building, Niels Bohrweg 1, 2333 CA Leiden)
Time: 15.30 - 17.15
ECTS: 6

Description:

The course Databases & Data Mining consists of a series of lectures in which advanced database and data mining techniques will be discussed, with applications to bioinformatics.

Grading:

There will be 2 database and 2 data mining assignments, i.e., 4 assignments in total, and a final exam (open book). The final grade will be based on a weighted average of the grades obtained for assignments P1, P2, P3, P4 and the Exam (E >5):

Final Grade = (0.5*P1 + P2 + 0.5*P3 + P4 + 3*E)/6.

DBDM: "Registration"

Please send an email

To: s.manegold@liacs.leidenuniv.nl
Subject: [DBDM-2018] Registration

containing the following information:

- *Your full name*
- *Your email address*
- *Your student ID*
- *Your affiliation (university)*
- *Your program / subject*

By Sunday 16 September 2018, 23:59 CEST.

DBDM: (tentative) Schedule

Date	Room	Subject (tentative)	Topic & Lecturer
11-09	312	Introduction	
18-09	312	Database Techology	Databases and Data Management for Data Mining
25-09	312	Database Techology	
02-10	312	Data Preprocessing	Stefan Manegold
09-10	*312*	*No class*	
16-10	312	Data Warehousing and OLAP	
23-10	312	Data Cube Technology	
30-10	312	Basic Data Mining Algorithms I	Data Mining Techniques and Applications
06-11	312	Basic Data Mining Algorithms II	
13-11	312	Advanced Data Mining Algorithms	Erwin Bakker
20-11	312	Mining in Bio-Data	
27-11	312	Graph Mining I	
04-12	312	Graph Mining II	

DBDM: Assignments

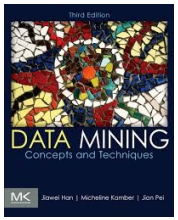
- 2 database assignments & 2 data mining assignments
- Will be announced individually during lectures and posted on website

DBDM: Exam

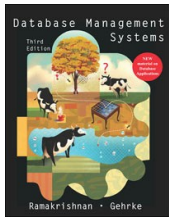
- **open book exam:** you can take with you your book, and printed course notes (slides). *No electronic equipment is allowed, though.*
- **Materials to be studied:**
  - All content covered and discussed during lectures (slides will be shared).
  - More to be announced.
- **Date:** *Monday, January 7, 2019*
- **Time:** *14:00 - 17:00*
- **Place:** *Room F104, Van Steenisgebouw, Einsteinweg 2, 2333 CC Leiden*

DBDM: Recommended Books

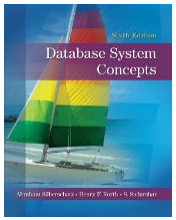
- **Data Mining:**
  - *J. Han, M. Kamber, J. Pei.* **Data Mining Concepts and Techniques (3rd Edition)**, Morgan Kaufman Publishers, July 2011 (ISBN 978-0123814791)



- **Database systems (e.g.):**
  - *Ramakrishnan, Gehrke*: **Database Management Systems (3rd International Edition)**, McGraw-Hill, 2003 (ISBN 0-07-246563-8)



  - *A. Silberschatz, H. F. Korth, S. Sudarshan:* **Database System Concepts (6th Edition),** McGraw-Hill, 2010 (ISBN 0-07-352332-1)



DBDM: "Registration"

Please send an email

To: s.manegold@liacs.leidenuniv.nl
Subject: [DBDM-2018] Registration

containing the following information:

- *Your full name*
- *Your email address*
- *Your student ID*
- *Your affiliation (university)*
- *Your program / subject*

By Sunday 16 September 2018, 23:59 CEST.

Databases & Data Mining

Stefan Manegold

Group leader Database Architectures
Centrum Wiskunde & Informatica (CWI)
Amsterdam
http://homepages.cwi.nl/~manegold/

http://www.monetdb.org/

Prof. Data Management (0.2 fte)
LIACS & LCDS
Faculty of Science, Leiden University

https://homepages.cwi.nl/~manegold/DBDM/

http://liacs.leidenuniv.nl/~bakkerem2/dbdm/

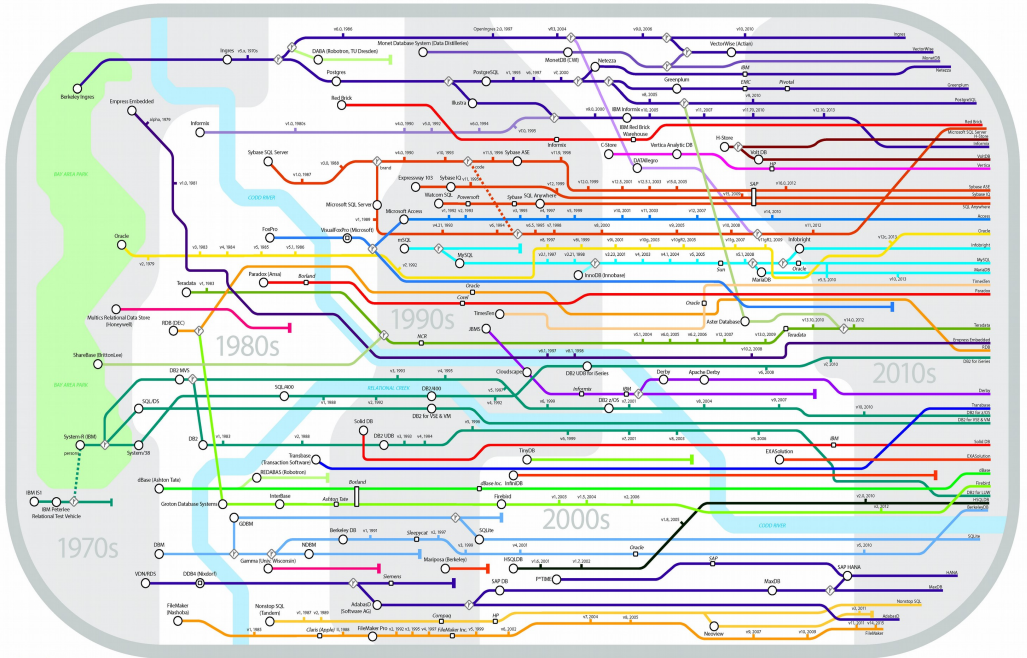
Data

Data Management

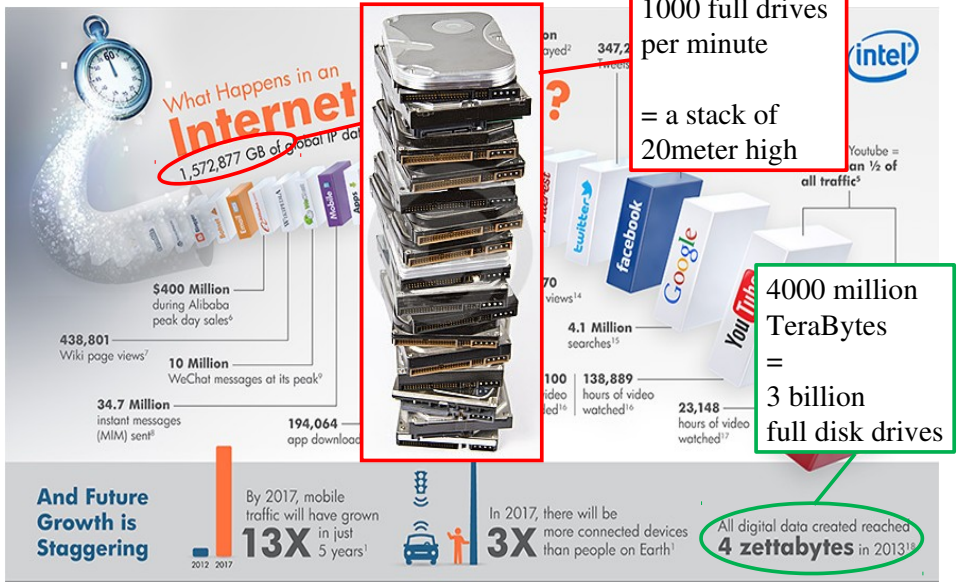
Database

Data Mining

Genealogy of Relational Database Management Systems



The age of Big Data



1500TB/min =
1000 full drives
per minute

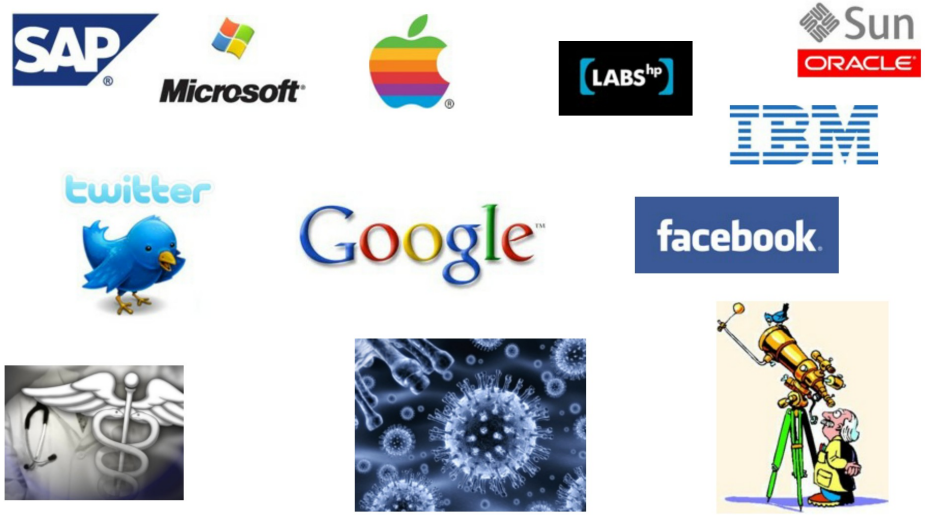
= a stack of
20meter high

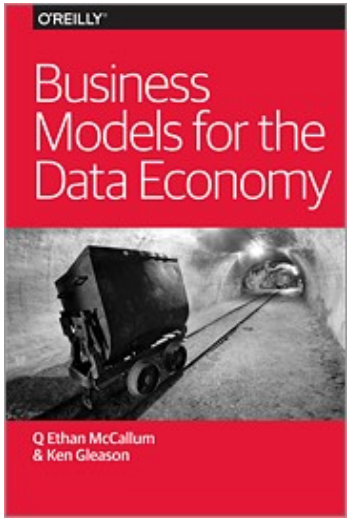
4000 million
TeraBytes
=
3 billion
full disk drives

CWI Database Architectures

Universiteit Leiden

The Data Economy





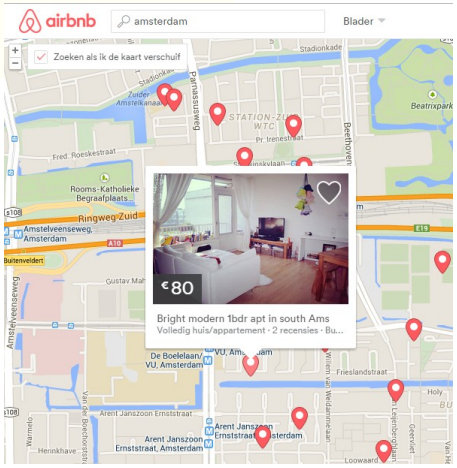
CWI Database Architectures

Universiteit Leiden

Disruptions by the Data Economy

UBER

EVERYONE'S PRIVATE DRIVER™



DBDM: Selected Challenges

GIS (LIDAR):

Massive point clouds: 640 Billion (x,y,z) points / 15 TB

=> spatial joins between point cloud and polygons

netherlands eScience center

Logistics:

TOMTOM > 5 trillion (10^12) GPS points (grows with >60k points/sec)

COMMIT/

Seismology:

Koninklijk Nederlands Meteorologisch Instituut Ministerie van Infrastructuur en Milieu ~ 4 M files, ~ 500 GB (10x compressed)

=> Transparent data ingestion: *Data Vault*

Remote sensing:

DLR ~2 PB satellite image data

TELEIOS

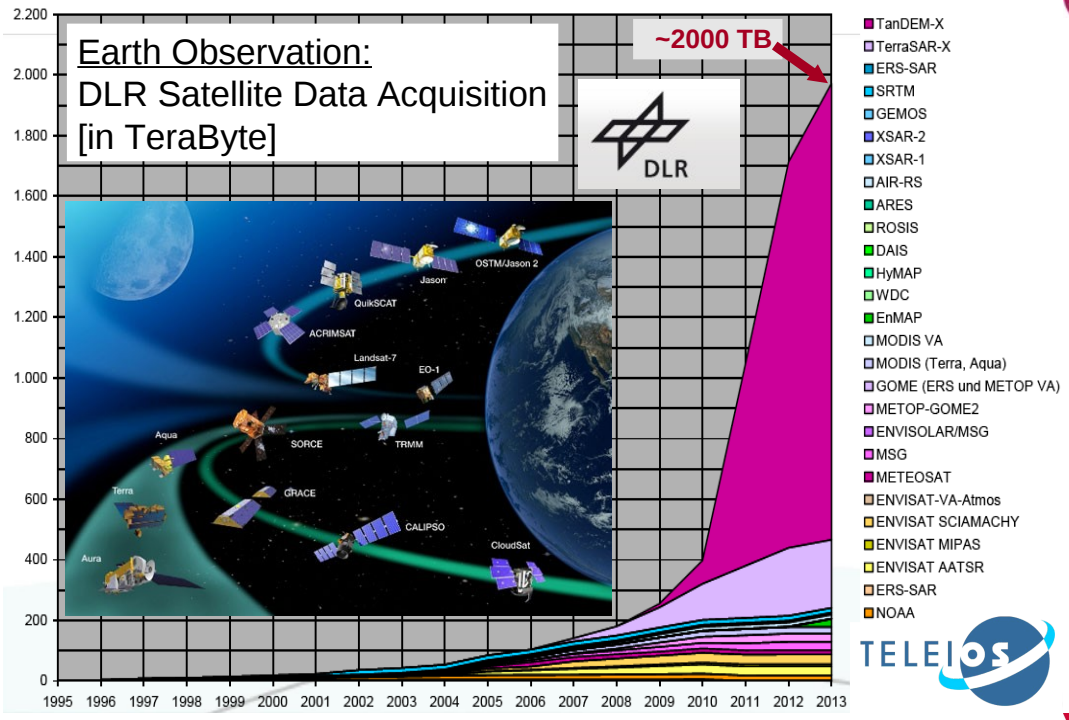
=> Array data processing: *SciQL*

Astronomy:

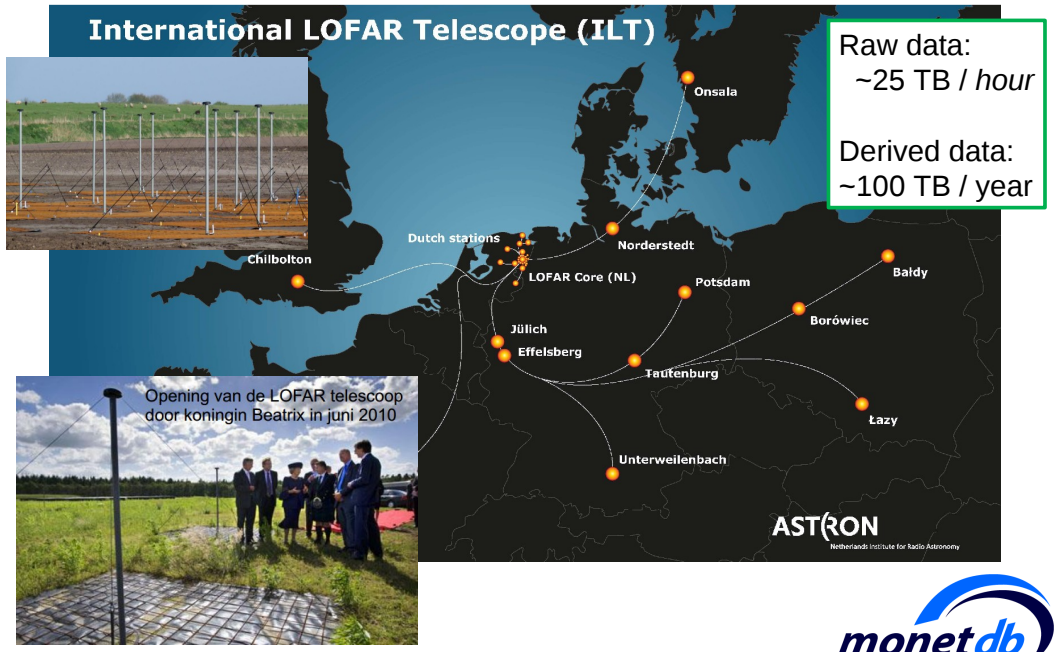
LOFAR Raw data: *25 TB / hour*; derived data: *100 TB / year*

=> Transient detection inside DBMS

DBDM: Earth Observation



LOFAR Low Frequency Array for Radio Asronomy



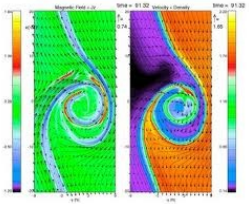
monetdb

Universiteit Leiden

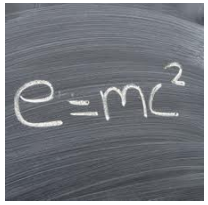
CWI Database Architectures

Data Disrupting Science:
Paradigm Shift in Scientific Research

collecting & analyzing data
data exploration
(eScience)

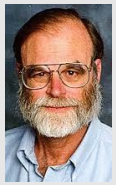


simulating
computational
3rd

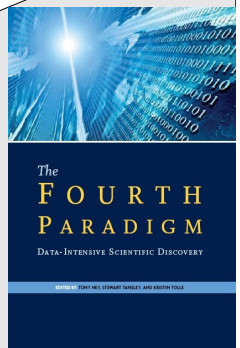


modeling
theoretical
2nd

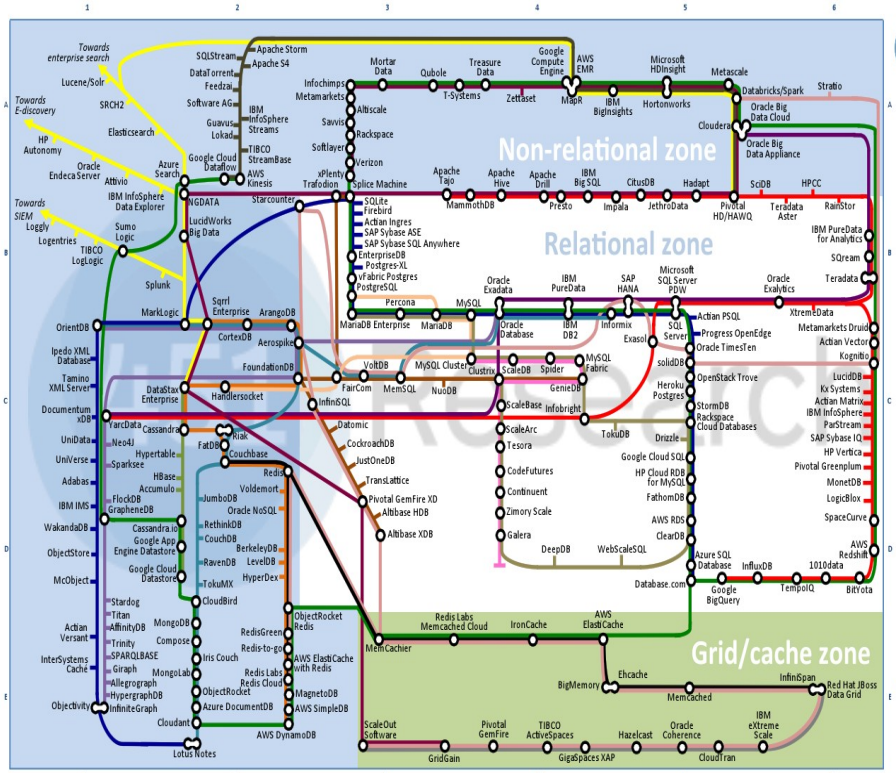
observing
empirical
1st



Jim Gray
(1944 - 2007)



Data Management & Data Mining



451 Research

Data Platforms Map

October 2014

Key:

- General purpose
- Specialist analytic
- -as-a-Service
- BigTables
- Graph
- Document
- Key value stores
- Key value direct access
- Hadoop
- MySQL ecosystem
- Advanced clustering/sharding
- New SQL databases
- Data caching
- Data grid
- Search
- Appliances
- In-memory
- Stream processing

https://451research.com/dashboard/dpa

BIG DATA LANDSCAPE, VERSION 3.0

Exited: Acquisition or IPO

Infrastructure

NoSQL Databases

redis, FOUNDATIONDB, DATASTAX, mongoDB, Couchbase, boshu, HYPERTABLE, aerospike, KEROSPIKE, CLOUDANT, amazon DynamoDB, OhmData, Neo4j, sones

Hadoop On Prem

HADAPT, cloudera, splice MACHINE, Zettaset, HSTREAMING, MAPR, amazon, Microsoft, Hortonworks, Pivotal, IBM InfoSphere BigInsights

Big Data

MORTAR, infochimps, bole, alticale, amazon web services, JETHRO DATA

NewSQL Databases

MarkLogic, TRANSLATTICE, RainStor, paradigm4, memsql, deepdb, citusdata, SkySQL, NUODB, clustrix, VoltDB, dugeo SQLFire

Cluster Service

LexisNexis, HPCC Systems, mesosphere, Acunu

MPP Data Stores

TERADATA, ParStream, InfiniDB, kognitio, VERTICA, NETEZZA, SQL Server, Pivotal, PARACCEL

Management / Monitoring

New Relic., metafor, StackIQ, tidemark, appnomic, AppFirst, ocean sync, DATADOG, boundary

Graph Databases

Neo4j, GIRAPH, aster data, InfiniteGraph

Data Transformation

TRIFACTA, Paxata, DataTamer, KALIDO, pentaho, TIKE iRON, syncsort

Security

DATAGUISE, codefortytwo, Stormpath, IMPERVA

Storage

Clearsofte, panasas, nimblestorage, Compuverde

Crowd-sourcing

microtask, CROWD COMPUTING, amazon mechanical turk, servio, mobileworks

App Dev

CONCURRENT, CONTINUITY, wibidata

Analytics

Analytics Platforms

databricks, QuantCell RESEARCH, PERVASIVE, guavus, Datameer, KARMASPHERE, collective[i], PRECOG, dataspora

For Business Analysts

STAT WING, CIRRO, TREPAREL, OrigamiLogic, ClearStory, DataGravity

Data Science Platforms / Tools

domino, Alpine, tonian, Sense, CONTINUUM ANALYTICS, MORTAR, plotly, yhat, MODE

Big Data

birst, JASPERSOFT, bime, pentaho, GoodData, SiSense, DOMO, platfora

Unstructured Data

BASIS, ATTIVIO, semantria, Crimson Hexagon, DIGITAL REASONING, Quid, Palantir

Data Visualization

tableau, ZoomData, visual.ly, Roambi, Chart.io, qlik, Quantum4D, ACTUATE, Kitenga, Looker, Ayasd, ISS, METRIC INSIGHTS, DataHero, GECKOBOARD

Machine Learning

SKYTREE, bigml, YOTTAMINE ANALYTICS, wise.io, context relevant

AI

IBM WATSON, reactor labs, vicarious

Social Analytics

simple reach, GNIP, bluefin, DATASIFT

bitly, synthesio, Datominr, Statistica, tracx, bottlenose

Location / People / Events

RADIUS, Fliptop, LOCATE, Locu, Recorded Future, PlaceIQ

Analytic Services

BIG DATA, THINK BIG, McKinsey&Company, accenture, OPERA, VALIANCE, DATA SCIENCE, Mu Sigma

Big Data Search

hp Autonomy, LucidWorks, ONTOLOGY

Statistical Computing

Prior Knowledge, REVOLUTION, SAS, MATLAB, SPSS

Log Analysis

splunk>, loggly, SUMO LOGIC, Kibana, sumologic

Crowd-sourced

kaggle, DataKind

Real-Time

METAMARKETS, MarketStream, amiato, causata

SMB

RJMetrics, retention, GoSquared, sumall, custora

Applications

Ad Optimization

aggregate knowledge, rocketfuel, TAPAD, MediaMath, ai Match, the trade desk, 33across, exelate, Data Xu, dstillery, m6d

Publisher Tools

Chartbeat, Yieldex, yieldbot

Marketing

LATTICE ENGINES, Sailthru, spinnakr, gainsight, RelateIQ, Kontera, Telkpart, persado, bloomreach, CLICKFOX, Pursway

Finance

Lenddo, BILLGUARD, cignifi, wonga, finance, LendUp, KENSHO, OnDeck

Human Resources

evolv, entelo, gild

Legal

JUDICATA, Lex Machina, RAVEL

Government / Regulator

mark43, enigma, Socrata, FiscalNote, Prism, PREDPOL

Security

SCINFYO, sift science, FORTSCALE, feedzai

Education / Learning

KNEWTON, declara, PANORAMA, Clever

Health

Recombine, 23andMe, Ginger.io, uBiome, FLATIRON, Counsyl

Industries

tubular, OPOWER, NEXT BIG SOUND, SIGHT MACHINE, THE CLIMATE CORPORATION, Fire

Cross Infrastructure / Analytics

SAP, sas, IBM, Google, Microsoft, vmware, amazon, hadoop, talend, TERADATA, hp Autonomy, NetApp

Open Source

Frameworks

Hadoop YARN, Spark, HOD FS

Query / Data Flow

Cascading, PIG

Data Access

Cassandra, SciDB, ORACLE, mongoDB, riak, HBase, CouchDB, Sqoop

Coordination / Workflow

ZooKeeper, talend

Real-Time

Storm

Stat Tools

SciPy

Machine Learning

Mahout

Cloud Deploy

Search

elasticsearch, Solr, LUCENE

Data Sources

Data Mkts

Windows Azure Marketplace, bluekai, knoema, DataMarket, factual

Data Sources

DATA.GOV, premise, YODLEE, VALIDIC, xignite, plaid, quandl, STANDARD TREASURY, human/api

Sensor Data

kinsa, STREETLINE, fitbit, RunKeeper, SKYCATCH, JAWBONE, LUMASENSE TECHNOLOGIES, Withings, BASIS, estimote

Incubators & School

zipfian, GA, INSIGHT, DataElite