

Data Analytics Infrastructure

Data Science SG
Nov 2015 Meetup

Le Nguyen The Dat
@lenguyenthedat



Backgrounds

ZALORA Group (2013 – 2014)

- Biggest online fashion retails in South East Asia
- Data Infrastructure & Data Science



Backgrounds

Commercialize.TV (2015 –)

- Multi Channel media network – focusing on China audiences
- Data Infrastructure & Insights

The screenshot displays the Commercialize.TV website interface. At the top is a dark navigation bar with the CTV logo and links for ABOUT US, PROGRAMMING, IMPACT FOR BRANDS, TALENTS, OUR PARTNERS, and CONTACT US. Below the navigation bar are three main content categories, each with a representative image and a description:

- KIDS**: Image of children playing. Description: "For your little ones – and you! On our kids network, you'll find everything you need to keep your kids entertained and inspired. From reviews, new animation blockbusters and fun, do-it-yourself projects – these are just some of the engaging content that will keep your child's curious mind captivated and inspired."
- FASHION**: Image of models on a runway. Description: "The ultimate curation of professional fashion show content from our leading industry partners alongside news, gossip, celebrity and hand-chosen emerging creator content. From the runways of New York to Paris, Milan and Shanghai we have exclusive behind the scenes access. We have rights to exclusive shows and simultaneously foster up-and-coming online fashion & beauty talent. The perfect mix of aspirational and accessible content from the East and the West."
- ACTION & ADVENTURE**: Image of a person in a rugged, outdoor setting. Description: "Thrilling adrenaline-pumping achievements in the great outdoors. A look-away global mix of athletes alongside first-hand experiences that breathe their passion. Snow sports, extreme sports, sublime immersion in the world's most breathtaking landscapes."

At the bottom of the page is a horizontal list of category icons and labels: LUXURY, FASHION, ACTION & ADVENTURE, TRAVEL, FOOD, ENTERTAINMENT, and KIDS.

Challenges

No central data source:

- Data stored in multiple locations
- Unclear ownership

Data definition and quality:

- Little to none documentation
- Different formula, rules owned by different departments
- Always dirty no matter what

Reporting – Descriptive analytics:

- Immediate needs, automations
- Important to do it right (and quick!)



Data Warehouse

Database Technologies

SQL – Relational Databases:

- MySQL, PostgreSQL
- MS SQL Server, Oracle SQL

NoSQL:

- Redis
- Cassandra
- MongoDB
- DynamoDB (AWS)
- RethinkDB

Map Reduce ecosystem:

- Hadoop: HDFS – Pig – Hive – Hbase
- Spark: RDD – Spork – Shark (Spark SQL) – Hbase-Spark

Massively Parallel Processing (MPP):

- Vertica (HP) - Greenplum (EMC) – Netezza (IBM)
- ParAccel (Amazon Redshift)



Database Technologies

SQL – Relational Databases: ✓

- MySQL, PostgreSQL
- MS SQL Server, Oracle SQL

NoSQL: (?)

- Redis ✗
- Cassandra ✓
- MongoDB ✗
- DynamoDB (AWS) ✗
- Neo4j ✗

Map Reduce ecosystem: ✓

- Hadoop: HDFS – Pig – Hive – Hbase
- Spark: RDD – Spork – Shark (Spark SQL) – Hbase-Spark

Massively Parallel Processing (MPP): ✓

- Vertica (HP) - Greenplum (EMC) – Netezza (IBM)
- ParAccel (Amazon Redshift)

Data Warehouse

Amazon Redshift

aws.amazon.com/redshift

- Cloud-based, Fully managed
- SQL (PostgreSQL 8.0.2)
- On-demand (\$2000/year)
- Scalable (Petabyte-scale)
- FAST! amplab.cs.berkeley.edu/benchmark/

All in ONE place!

- Product information
- Customer information
- Tracking data
- External data sources
(Social Media, 3rd Party datasets)

CONTAINS

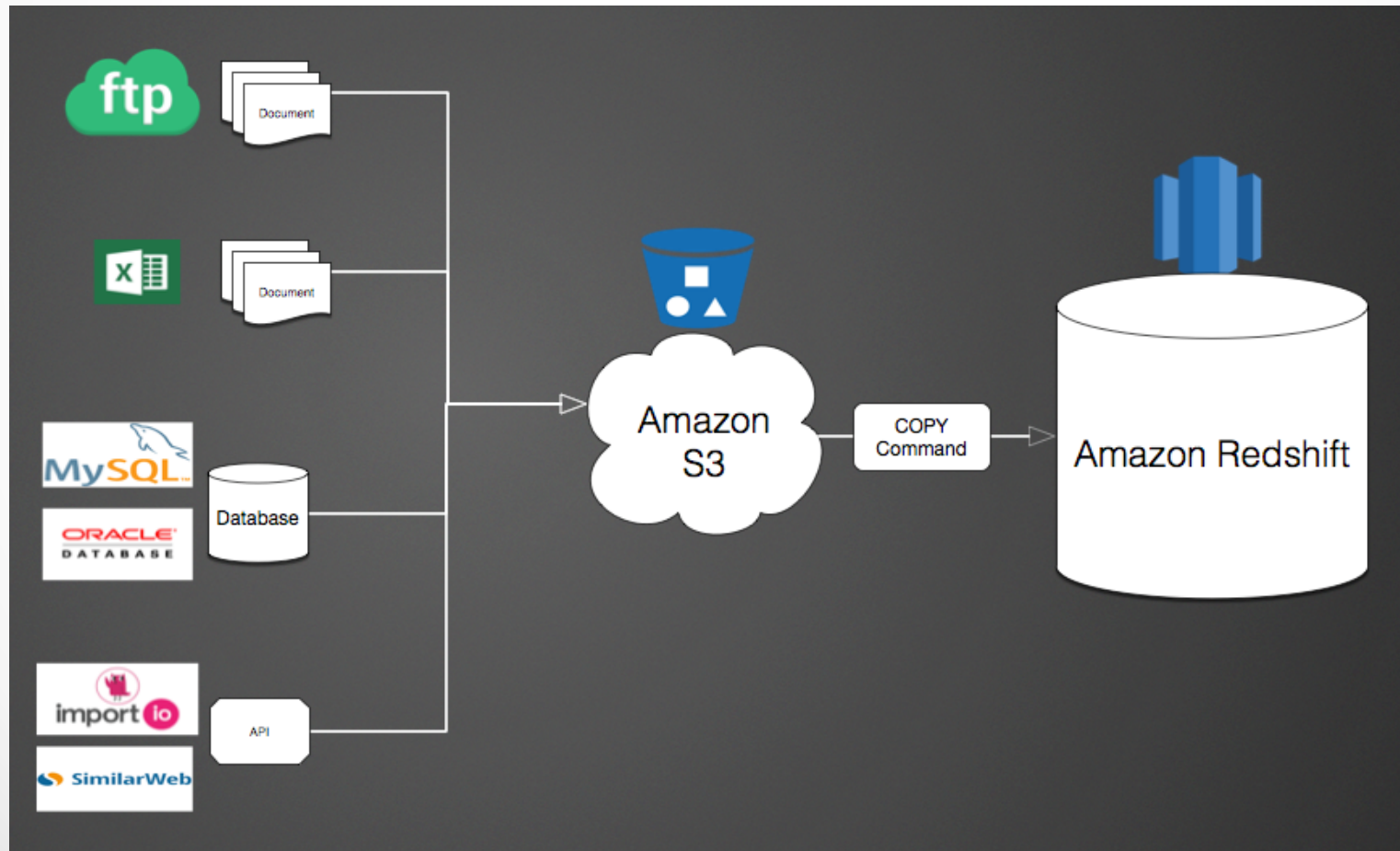


Extract-Transform-Load (ETL)

ETL

Amazon Redshift's COPY command.

docs.aws.amazon.com/redshift/latest/dg/r_COPY.html



ETL

Custom made:

- Simple bash script, python, SQL
- Use cases:
 - Scrappers
 - Excel / CSV imports

Data Pipeline Frameworks:

- Large scale, more complicated
- Examples:
 - Spotify's Luigi – github.com/spotify/luigi
 - Yelp's Mycroft – github.com/Yelp/mycroft

3rd Party Services:

- aws.amazon.com/redshift/partners
 - Flydata
 - Rjmetrics

Applications

Self-services Dashboards

Intermediate users

- SQL / Excel / WYSIWYG Tools
- Re:dash – www.redash.io
 - Open source: github.com/getredash/redash
 - Try it out: demo.redash.io
 - Self-manage & deployment:
 - Docker
 - Pre-baked AMI (Amazon Web Services)
 - Google Cloud Images
 - Supports lots of database types (Redshift, MySQL, PostgreSQL, Big Query, MongoDB...)
 - Users need to know SQL
 - Web-based, collaborative work type

Demo: re:dash data sources usage

<http://demo.redash.io/queries/756>

Execute Format SQL

Fork Save

Search schema...

```
1 SELECT data_sources.name, date(queries.created_at), count(*)
2 FROM queries
3 LEFT JOIN data_sources ON data_sources.id = queries.data_source_id
4 WHERE date(queries.created_at) > CURRENT_DATE - INTERVAL '1 months'
5 GROUP BY data_sources.name, date(queries.created_at)
```

cohort2
cohort_example
dashboards
data_sources
events
queries
query_results
visualizations
widgets

Created By Le Nguyen The Dat

Last update 11 minutes ago

Runtime 0s

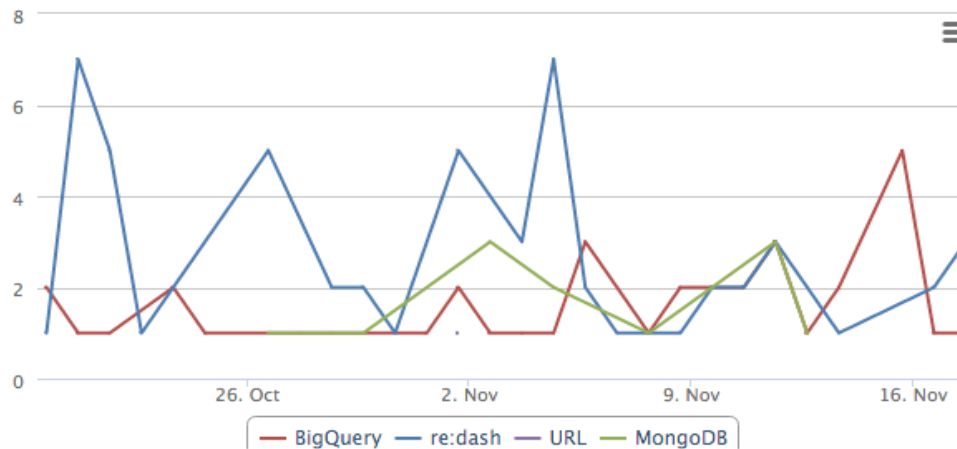
Rows 52

Refresh Schedule Every 24h

Data Source re:dash

Download Dataset

Table Pivot Table Line Chart x + New Visualization



Demo: NYC Taxis Tip Amounts

<http://demo.redash.io/queries/753>

Execute Format SQL

Fork Save

```
1 SELECT INTEGER(ROUND(FLOAT(tip_amount) / FLOAT(fare_amount) * 100)) tip_pct,  
2     count(*) trips  
3 FROM [833682135931:nyctaxi.trip_fare]  
4 WHERE payment_type='CRD'  
5     AND float(fare_amount) > 0.00  
6 GROUP BY 1  
7 ORDER BY 2 DESC LIMIT 20
```

Created By Le Nguyen The Dat

Last update 2 days ago

Runtime 3s

Rows 20

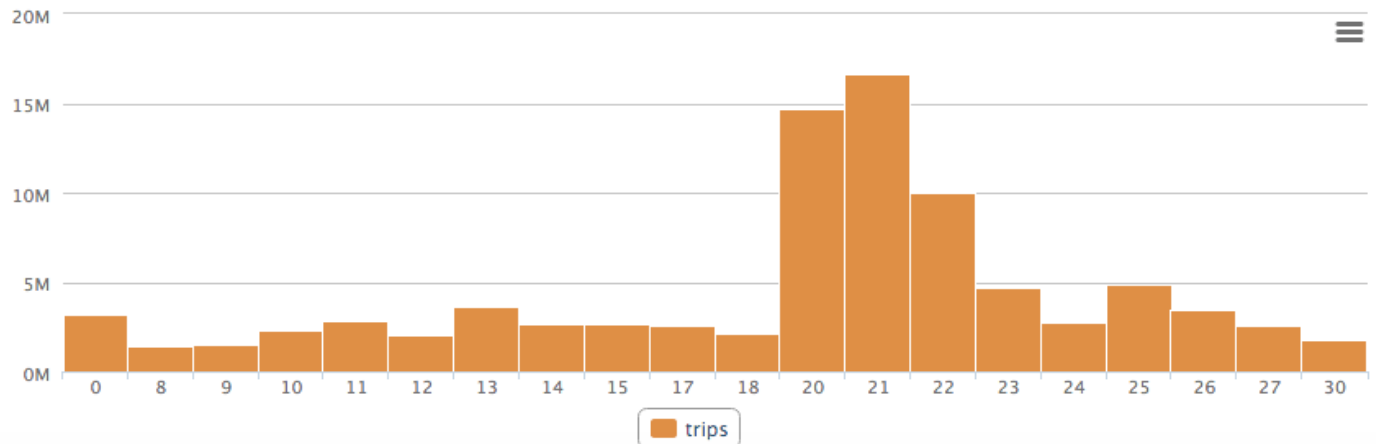
Refresh Schedule Never

Data Source BigQuery

Download Dataset



Table Pivot Table Pie Chart x Bar Chart x + New Visualization



Self-services Dashboards

Intermediate users

- SQL / Excel / WYSIWYG Tools
- Tableau – www.tableau.com
 - Licensed software (14 days trial)
 - Tableau Public (Free: public.tableau.com)
 - Self-host or Tableau host (fully managed)
 - Supports a lot more database types
 - Group, User management – customized access right
 - Drag & Drop software as well as web-based

Demo: Social media dashboard

Baidu → Import.io API → Data Warehouse → Tableau

[Share](#) • Remember my changes ▾ Edit

Date Range

Last 30 days ▾

Platform

Baidu ▾

Type

- ☐ (All)
- ☒ Top 10 Hot Celebrities
- ☐ Top 10 Hot Foreign Celebrities
- ☐ Top 10 Hot People
- ☐ Top 50 Searches

Measure Values

20,854  90,349

Name

-  胡歌
-  angelababy
-  鹿晗
-  赵丽颖
-  陈学冬
-  霍建华
-  王凯
-  王晓晨
-  杨洋
-  靳东
-  吴磊
-  喻虹渊
-  唐嫣
-  沈梦辰
-  毛晓彤
-  王丽坤
-  周渝民
-  于和伟
-  陈翔
-  朱亚文

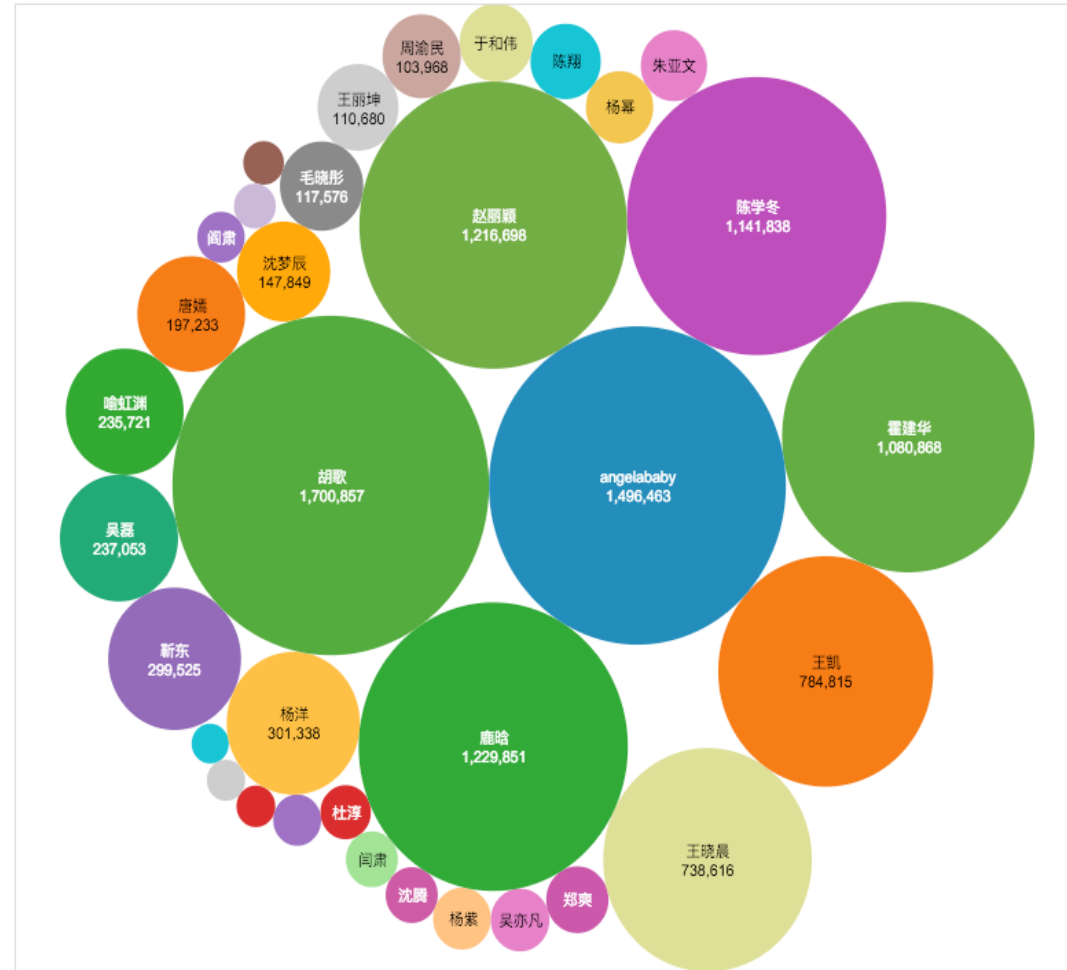
China Top Contents

At Date	Rank	Name	Views / Searches
2015-11-17	1	霍建华	62,877
	2	鹿晗	47,699
	3	胡歌	43,362
	4	朱亚文	41,379
	5	赵丽颖	40,034
	6	angelababy	37,357
	7	王凯	27,894
	8	陈学冬	26,225
	9	沈佳妮	25,513
	10	靳东	22,950
2015-11-16	1	鹿晗	78,012
	2	霍建华	65,572
	3	angelababy	60,244
	4	赵丽颖	44,258
	5	胡歌	42,216
	6	萧亚轩	38,649
	7	陈学冬	34,509
	8	王凯	27,032
	9	杨洋	25,463
	10	李晨	23,422
2015-11-15	1	霍建华	63,467
	2	鹿晗	47,454
	3	胡歌	45,248
	4	angelababy	43,145
	5	赵丽颖	39,758
	6	王凯	36,774
	7	朱亚文	33,114
	8	陈学冬	30,233
	9	靳东	24,825
	10	马伊P	24,808
2015-11-14	1	喻虹渊	86,006
	2	霍建华	44,149

 Download

China Top Contents Chart



Demo: Market Research - video platform performance

SimilarWeb API → Data Warehouse → Tableau

[Share](#) Remember my changes Edit



Download

Date Range

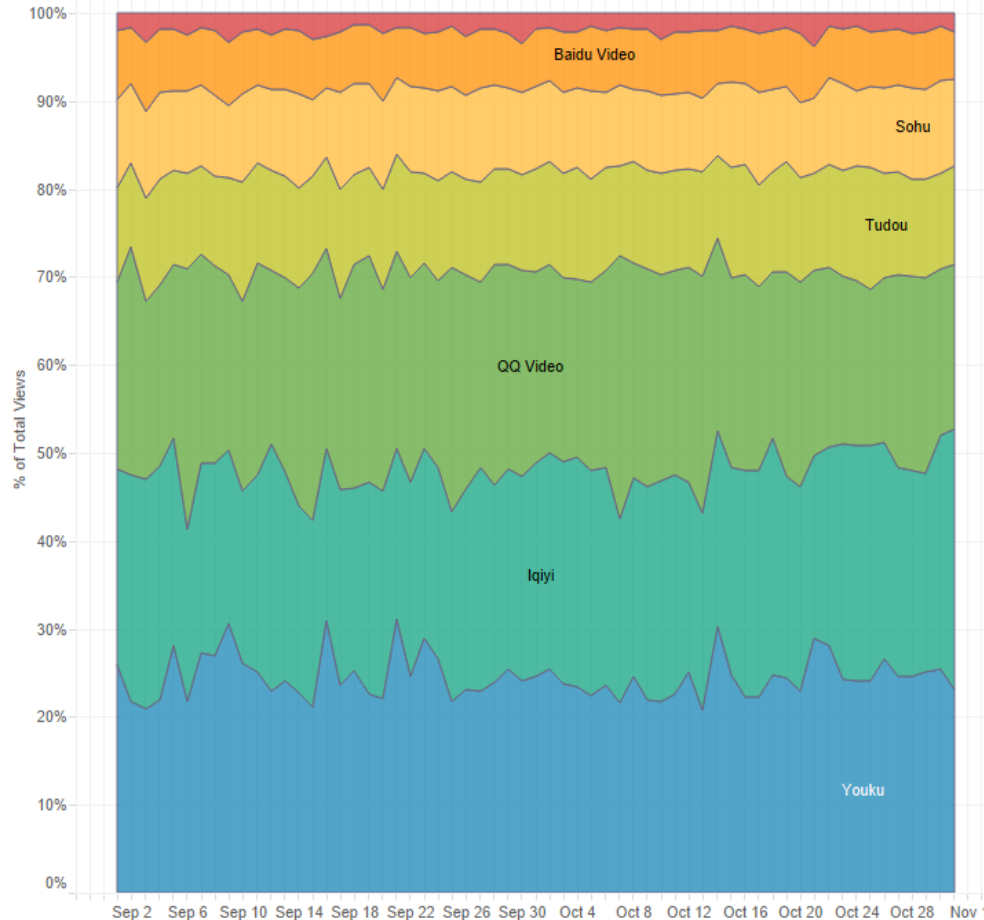
Last 3 months

Market Research Chart % (SimilarWeb)

% of Total Views
0.99% 56.99%

Domain

Sina Video
Baidu Video
Sohu
Tudou
QQ Video
Iqiyi
Youku

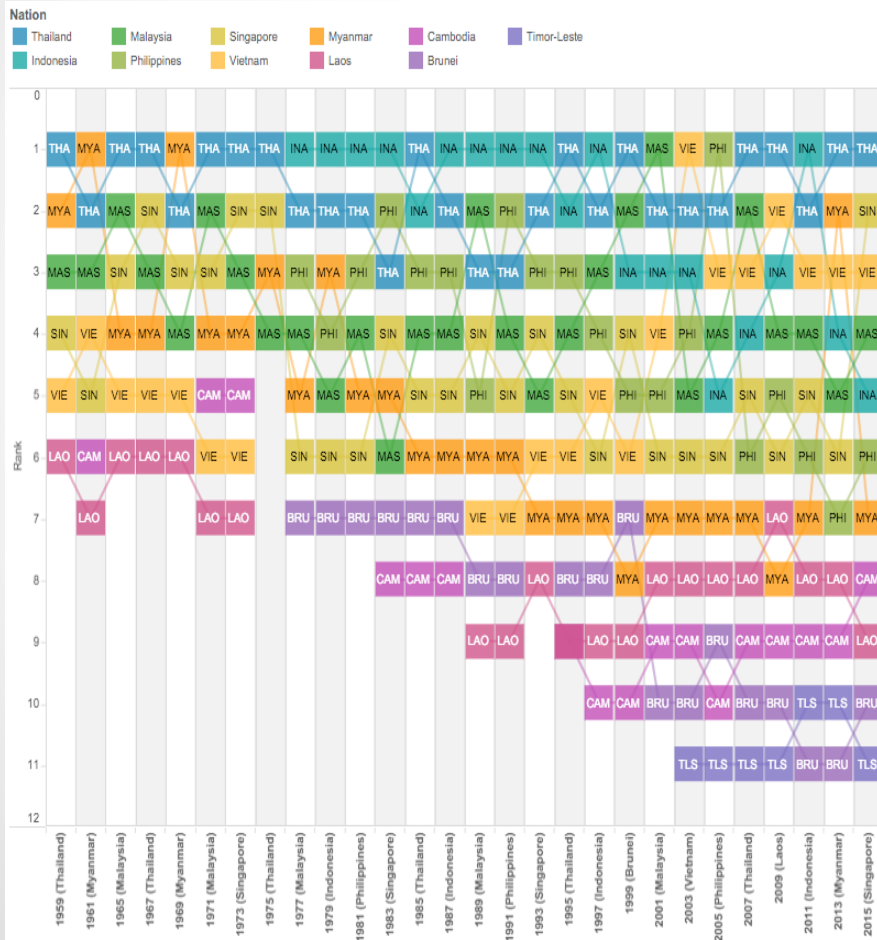


Market Research Heatmap % (SimilarWeb)

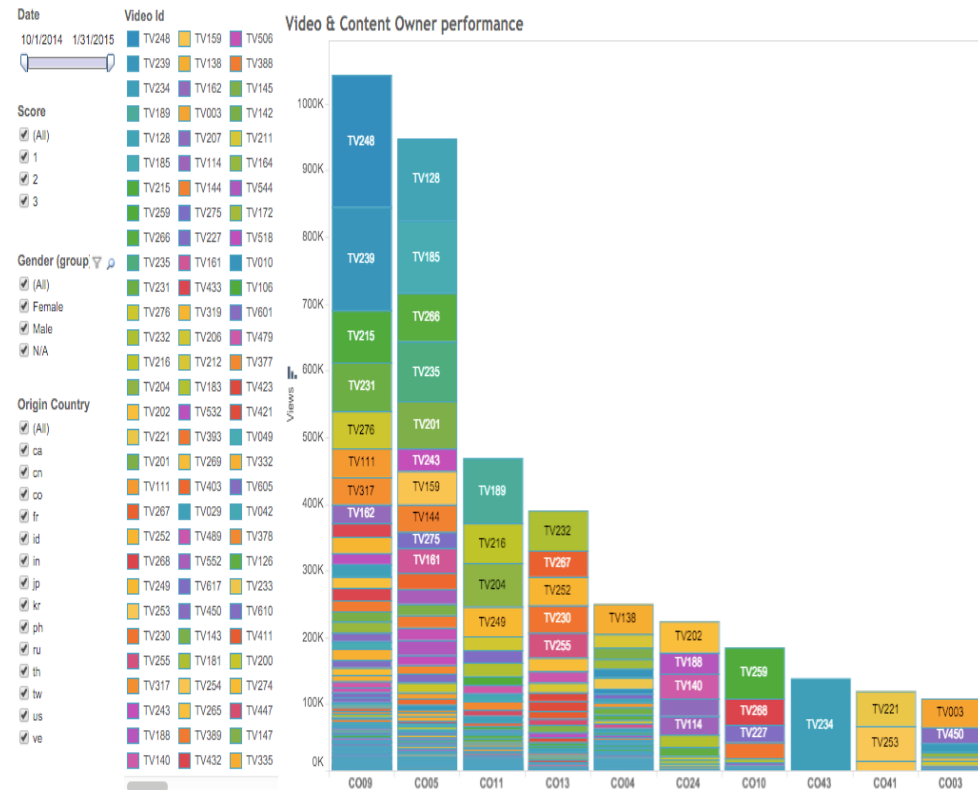
	Youku	Iqiyi	QQ Video	Tudou	Sohu	Baidu Video	Sina Video
2014-01-01	30.09%	8.22%	8.43%	14.89%	9.36%	23.63%	5.38%
2014-01-02	32.12%	9.33%	7.24%	15.82%	9.37%	21.15%	4.97%
2014-01-03	29.82%	9.12%	9.04%	14.72%	10.26%	22.34%	4.70%
2014-01-04	25.81%	8.84%	7.80%	14.12%	8.63%	30.61%	4.19%
2014-01-05	30.22%	9.45%	7.66%	14.69%	10.72%	23.00%	4.26%
2014-01-06	30.71%	9.95%	7.80%	14.86%	10.36%	21.75%	4.58%
2014-01-07	30.58%	9.73%	10.07%	14.34%	10.17%	20.38%	4.73%
2014-01-08	28.36%	9.45%	10.92%	15.97%	10.57%	18.81%	5.91%
2014-01-09	28.44%	10.05%	8.48%	16.09%	11.28%	20.44%	5.21%
2014-01-10	29.71%	9.80%	9.15%	14.27%	10.50%	21.34%	5.24%
2014-01-11	26.97%	8.78%	13.74%	12.87%	10.68%	21.69%	5.26%
2014-01-12	29.60%	10.64%	7.30%	15.65%	10.97%	22.04%	3.80%
2014-01-13	30.99%	9.70%	8.90%	14.12%	10.09%	20.42%	5.77%
2014-01-14	33.49%	9.33%	9.05%	14.53%	10.52%	18.32%	4.75%
2014-01-15	28.35%	10.13%	9.92%	13.26%	11.01%	20.88%	6.45%
2014-01-16	31.48%	10.27%	7.96%	14.42%	9.68%	19.36%	6.83%
2014-01-17	27.37%	12.81%	8.59%	13.14%	9.65%	24.07%	4.36%
2014-01-18	28.57%	14.58%	6.75%	13.20%	10.91%	21.27%	4.72%
2014-01-19	27.01%	14.24%	7.75%	13.37%	10.86%	22.40%	4.38%
2014-01-20	27.16%	15.84%	6.77%	14.38%	8.29%	23.32%	4.25%
2014-01-21	27.99%	15.63%	7.52%	13.84%	9.38%	20.86%	4.78%
2014-01-22	27.80%	16.13%	7.07%	14.34%	9.84%	20.13%	4.69%
2014-01-23	27.69%	17.05%	7.58%	13.97%	10.08%	18.06%	5.57%
2014-01-24	26.65%	15.79%	7.31%	13.29%	9.90%	22.29%	4.78%
2014-01-25	26.15%	16.01%	7.65%	12.62%	10.79%	22.44%	4.33%
2014-01-26	28.16%	14.55%	6.87%	14.24%	9.59%	22.14%	4.44%
2014-01-27	27.79%	15.49%	6.89%	14.22%	10.26%	21.96%	3.39%
2014-01-28	27.90%	16.02%	8.33%	14.56%	9.18%	20.38%	3.64%
2014-01-29	27.32%	15.18%	7.88%	14.65%	10.29%	20.05%	4.63%
2014-01-30	27.18%	17.82%	9.05%	14.11%	10.00%	17.50%	4.34%
2014-01-31	27.08%	12.18%	8.03%	14.51%	11.94%	22.36%	3.90%

Others (Tableau Public):

- SEA Games Result history tiny.cc/seagames
- Rakuten – Viki data challenge tiny.cc/viki-viz



Viewing Patterns	Origin Country - Language	Origin Country - Gender	Content Owners Performance	Origin Country - Genre	Videos Performance
------------------	---------------------------	-------------------------	----------------------------	------------------------	--------------------



Advanced applications

Advanced users

- Data Warehouse connection (JDBC - PostgreSQL)
- Automated, highly customized reports.
- Data Science:
 - Recommendation engine
 - Predictive modeling
 - Classifications

Advanced applications

Internal reporting tool

Data Warehouse → SQL, Python (Django), JS → Product-Finder

Product Finder [Black dress | SKU or ID | Tiffany | Atmosphere]

find

☒ SKU ☒ SKU Supplier Config ☒ Brand Name ☒ Product Name

Country - Supplier Type - Buying Brand Department - Season - Buying Planning Category - Sub Category -

Results per page 10 20 50 100

1 2 3 4 5 6 7 8 9 10 ...

Country	SKU Info	Stock (OMS)	Stock (BOB)	Warning!	# Net Sale	Rate of Sale	# Revenue	# Margin	Webtrekk Data	Warning!	
AU	[ONLINE]  AUD \$79.95 NOW AUD \$1.98 SKU AT048AA66DKN SKU Supplier Config AWS13KD261 ID Catalog Config 125333 Product Name Tiffany Bodycon Dress Color Family black Brand AtmosHere Brand Type Private Label Sub Category Dresses Buying Planning Category Wapp Buying Brand Department N/A Season Sum14Main Supplier Type Main Store				414	Last 7 Days Sold Returned/Rejected Cancelled Last 30 Days Sold Returned/Rejected Cancelled Total Sold Returned/Rejected Cancelled	Per Day Per Week Per Month (Days Online)	Last 7 Days Intake After Mkd Trading Last 30 Days Intake After Mkd Trading Total Intake After Mkd Trading	# Margin	Last 7 Days Impressions Views Visits CR CTR Last 30 Days Impressions Views Visits CR CTR Total Impressions Views Visits CR CTR	

Product Info

Sales Info

Tracking Info

Advanced applications

Recommendation Engine

Data Warehouse → SQL, Python, Haskell → ZALORA Website

The image displays four examples of product recommendations on the Zalora website, arranged in a 2x2 grid. Each example shows a main product image on the left, a red arrow pointing to a 'Similar products:' section in the middle, and a list of four recommended products on the right. The recommended products are shown in a horizontal row with left and right navigation arrows. Each product image has a vertical strip of smaller thumbnail images on its left side.

- Top Left:** A woman wearing a brown sleeveless top. The 'Similar products:' section shows four different styles of sleeveless tops.
- Top Right:** A white and black high-heeled sandal. The 'Similar products:' section shows four different styles of high-heeled sandals.
- Bottom Left:** A man wearing a blue and white checkered short-sleeved shirt. The 'Similar products:' section shows four different styles of short-sleeved shirts.
- Bottom Right:** A man wearing khaki shorts. The 'Similar products:' section shows four different styles of shorts. A black circular badge on the main image says 'ENDS JUNE 10'. A black banner at the bottom of the main image says 'EXTRA 20% OFF CODE: MNG20'.

Conclusions

Team & Technology stack

- Small team of 1-4 programmers
- Amazon Web Services
 - *No upfront cost*
 - *Low maintenance*
 - *Scalability*
 - *Integrations*
- Shell Scripts, Python, Haskell, D3.js
- Unix, Open-source technologies



Takeaways

- (Good) data infrastructure is important:
 - Build it first (before you hire a data scientist!)
 - Build it right: stable – fast – scalable.
- There is no silver bullet:
 - Understand what you need
 - Always do more research
- Data infrastructure is NOT that hard!
 - Utilize existing, modern technologies
 - Avoid old, proprietary technology that were built for the 90s!

References

Engineering Blogs and Tutorials

- <https://blog.asana.com/2014/11/stable-accessible-data-infrastructure-startup/>
- https://medium.com/@samson_hu/building-analytics-at-500px-92e9a7005c83
- <https://engineering.pinterest.com/blog/powering-interactive-data-analysis-redshift>
- <http://engineering.ifttt.com/data/2015/10/14/data-infrastructure/>
- <https://blog.rjmetrics.com/2015/10/15/the-data-infrastructure-meta-analysis-how-top-engineering-organizations-built-their-big-data-stacks/>
- <https://www.youtube.com/watch?v=reQtXquDpzo>
- <https://www.periscopedata.com/amazon-redshift-guide>

Benchmarks

- <https://amplab.cs.berkeley.edu/benchmark/>
- <https://www.flydata.com/blog/with-amazon-redshift-ssd-querying-a-tb-of-data-took-less-than-10-seconds/>
- <https://www.flydata.com/blog/hive-and-redshift-a-brief-comparison/>

Thank you!