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

ZALORA

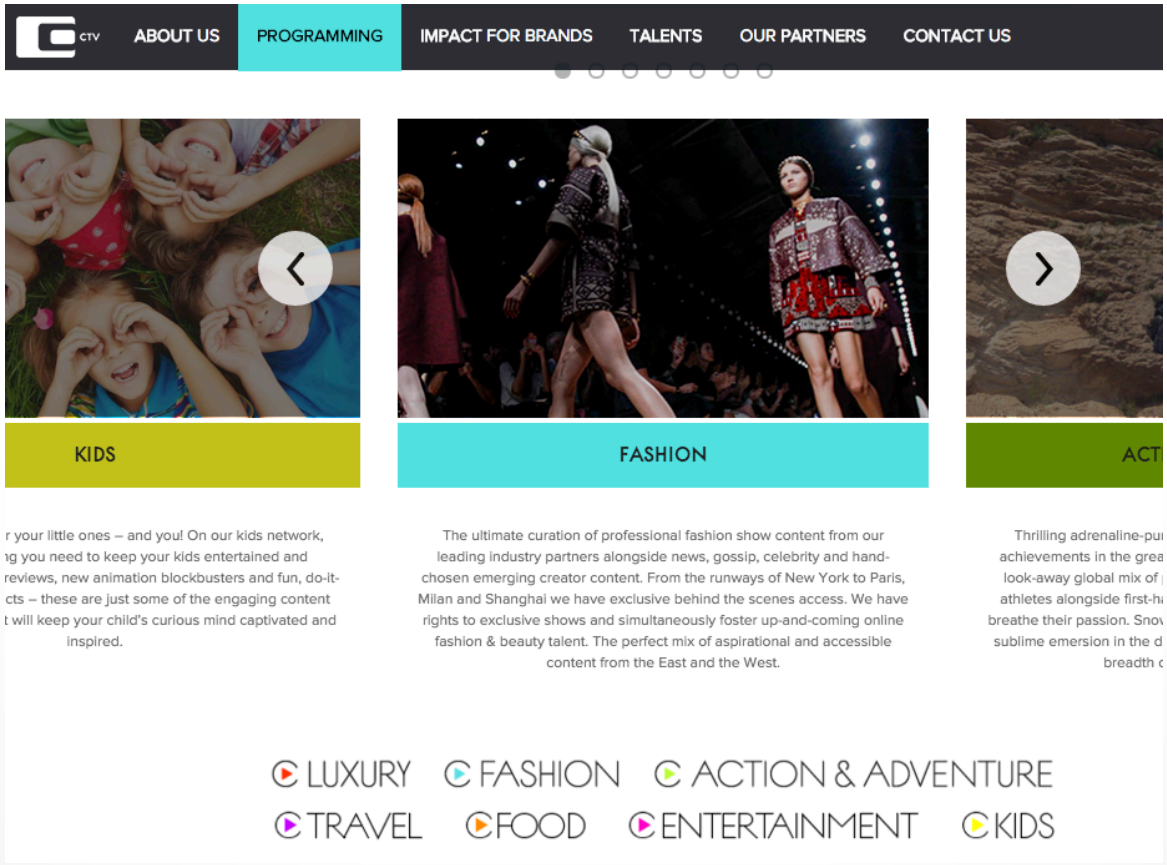
ASIA'S ONLINE FASHION DESTINATION

MALAYSIA	SINGAPORE
BRUNEI	HONG KONG
TAIWAN	PHILIPPINES
VIETNAM	THAILAND
INDONESIA	AUSTRALIA
NEW ZEALAND	

Backgrounds

Commercialize.TV (2015 –)

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



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!

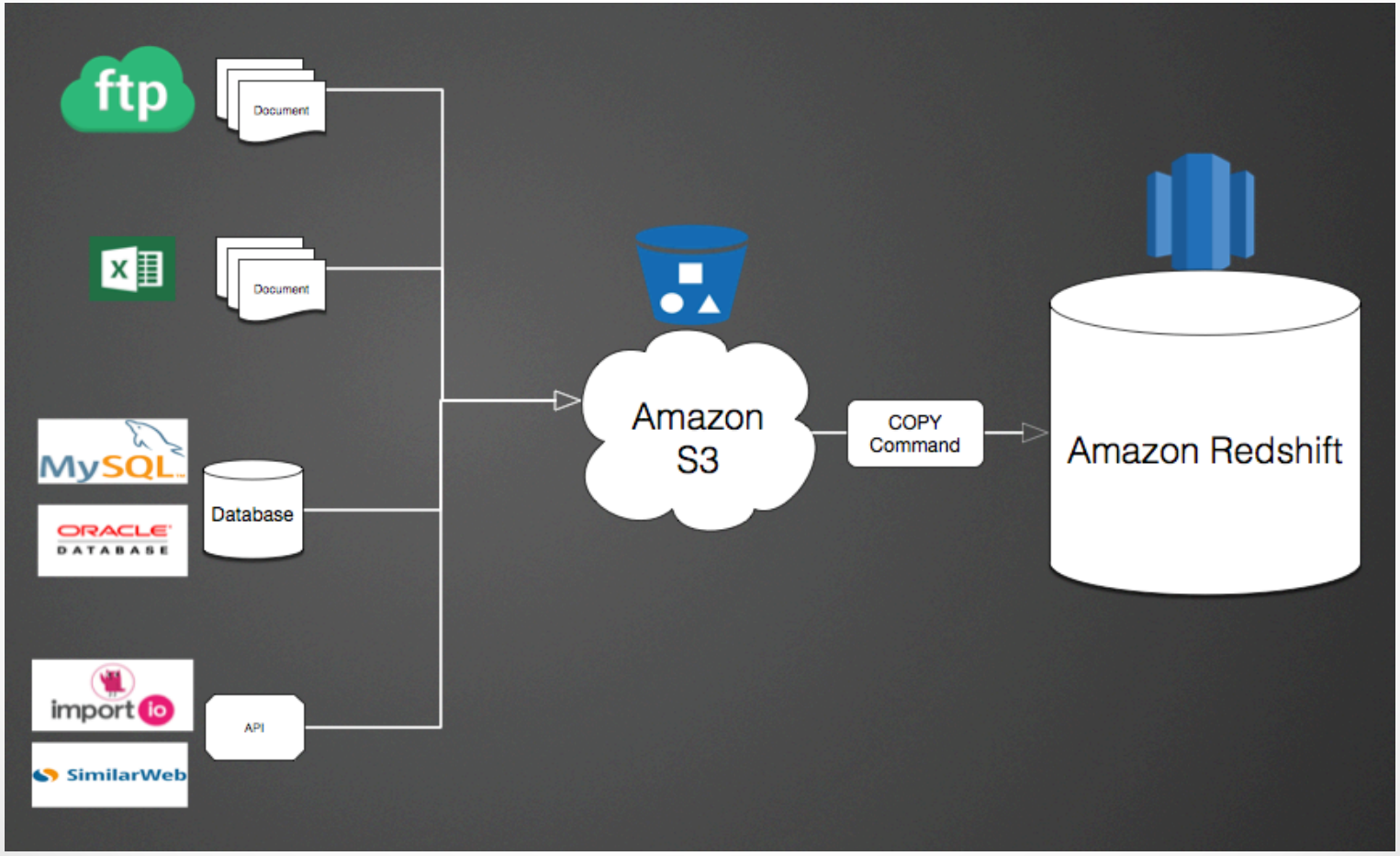
- Product information
- Customer information
- Tracking data
- External data sources

(Social Media, 3rd Party datasets)

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

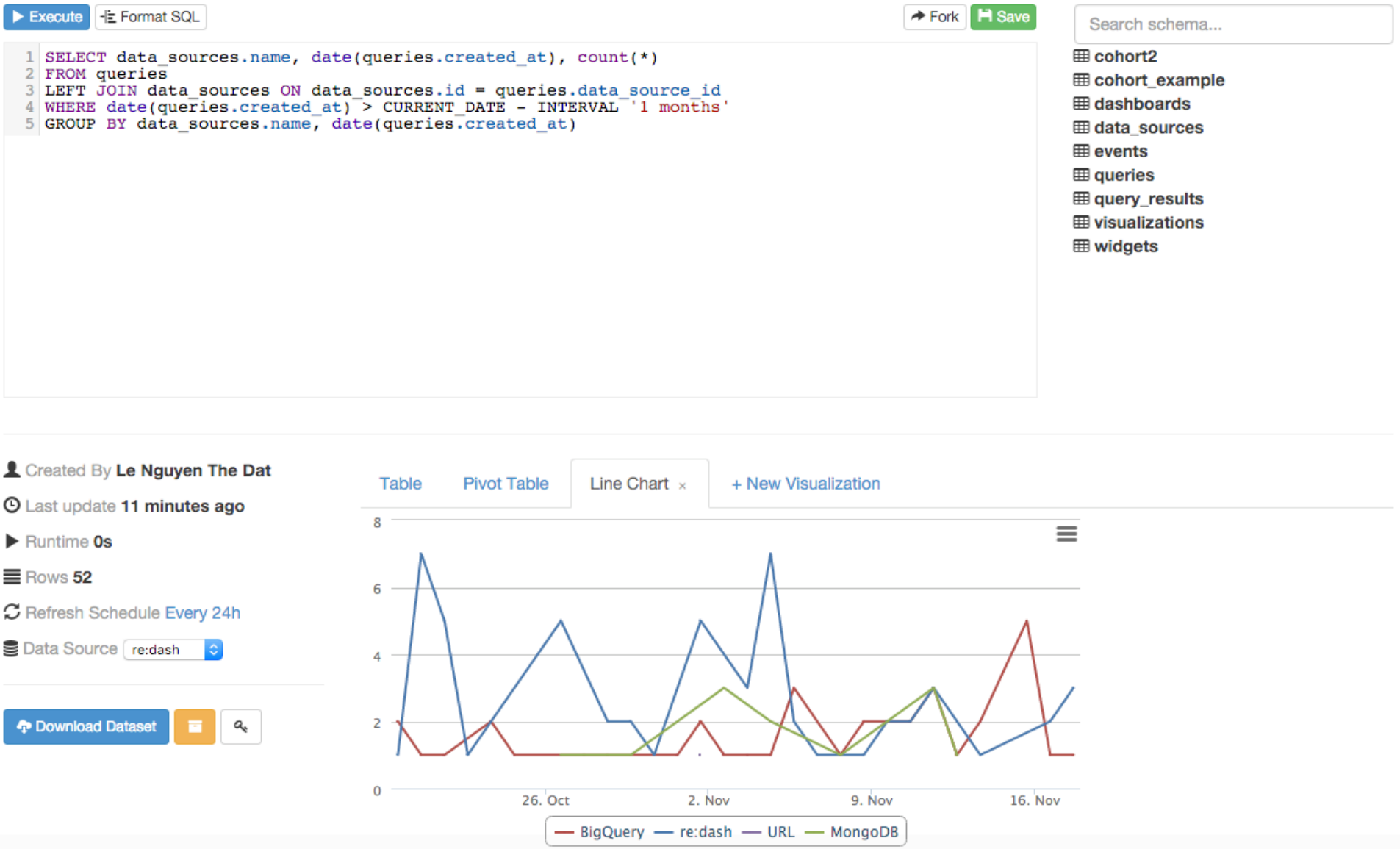
Self-services Dashboards

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



Demo: NYC Taxis Tip Amounts
http://demo.redash.io/queries/753

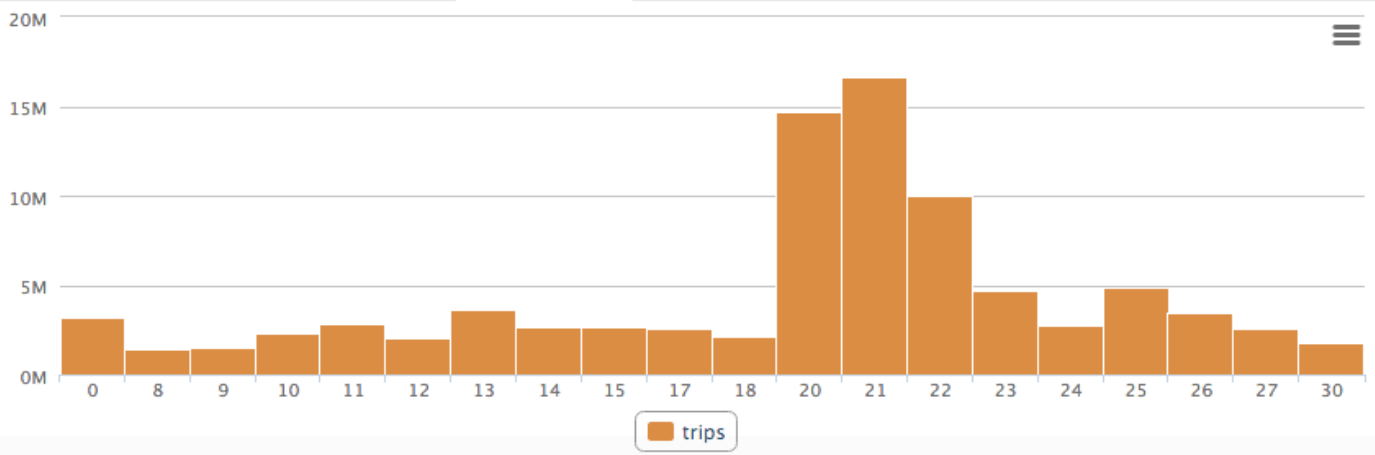
Execute | Format SQL | Fork | Save

```
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 × | Bar Chart × | + New Visualization



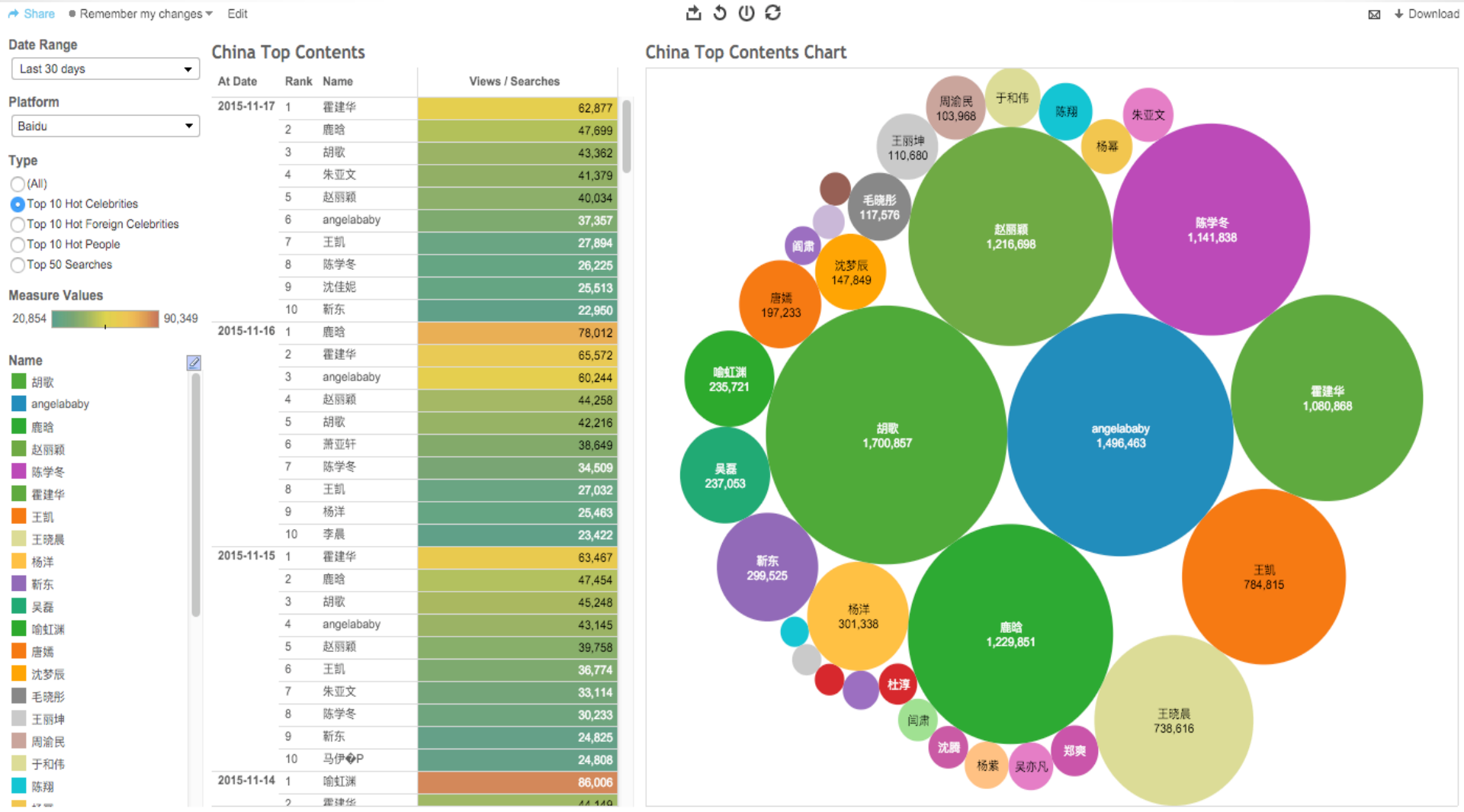
Self-services Dashboards

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

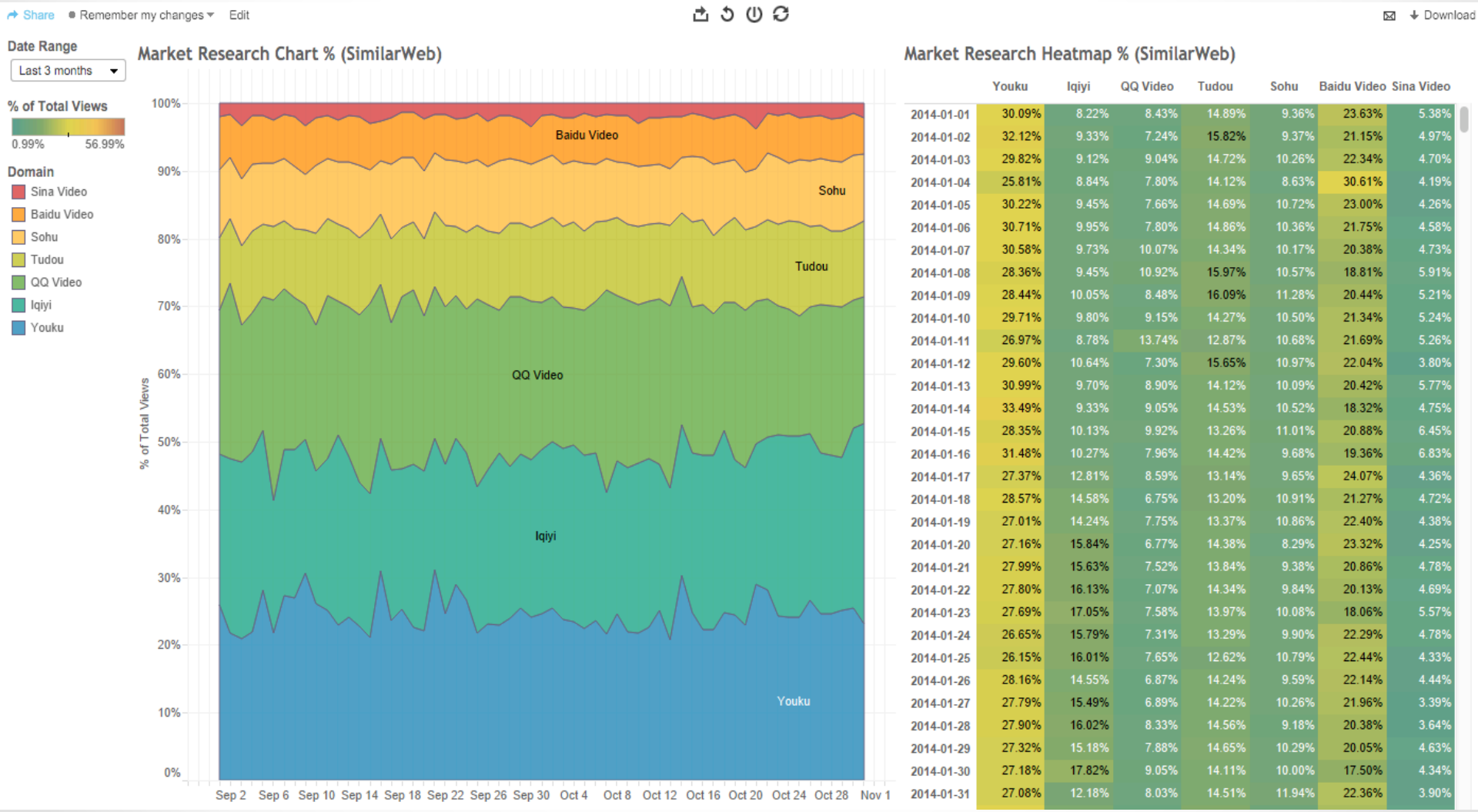
Demo: Social media dashboard

Baidu ➔ Import.io API ➔Data Warehouse ➔ Tableau



Demo: Market Research - video platform performance

SimilarWeb API ➔Data Warehouse ➔ Tableau

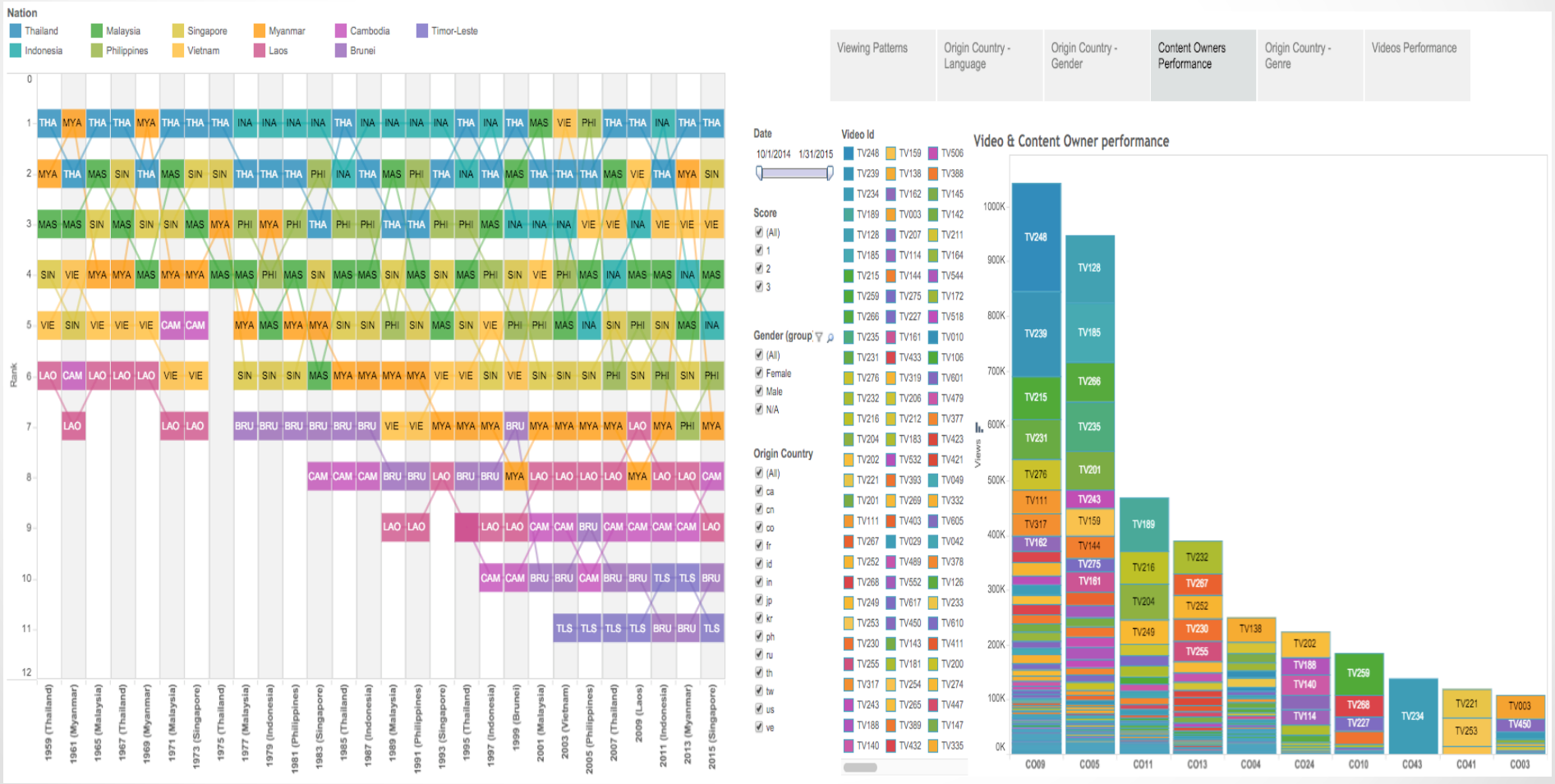


Market Research Heatmap % (SimilarWeb)

	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



Advanced applications

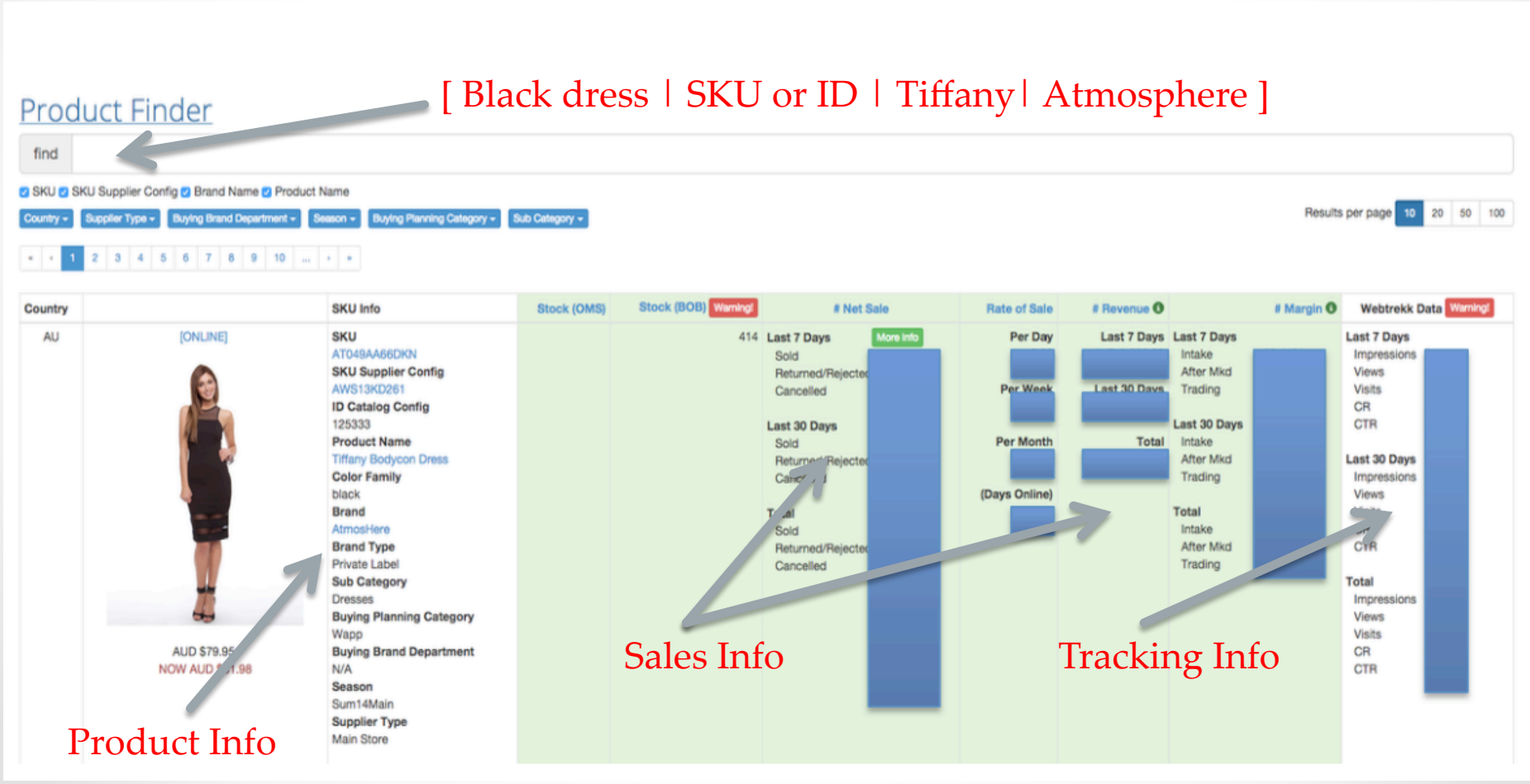
Advanced users

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

Advanced applications

Internal reporting tool

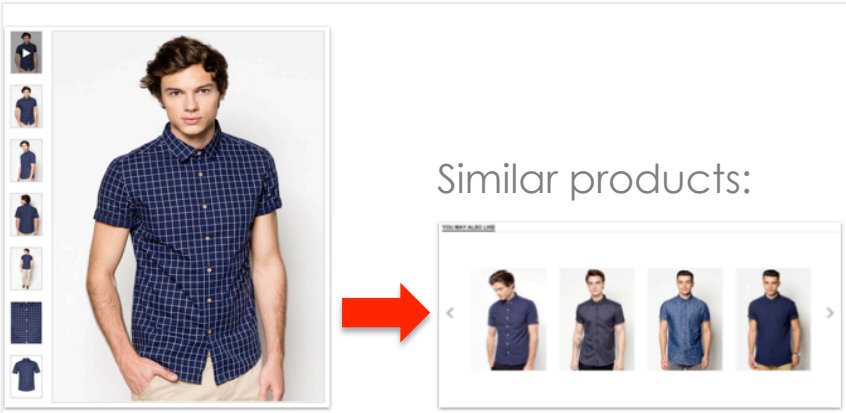
Data Warehouse ➔ SQL, Python (Django), JS ➔ Product-Finder



Advanced applications

Recommendation Engine

Data Warehouse ➔ SQL, Python, Haskell ➔ ZALORA Website



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!