



Building prediction models with Amazon Redshift and Amazon ML

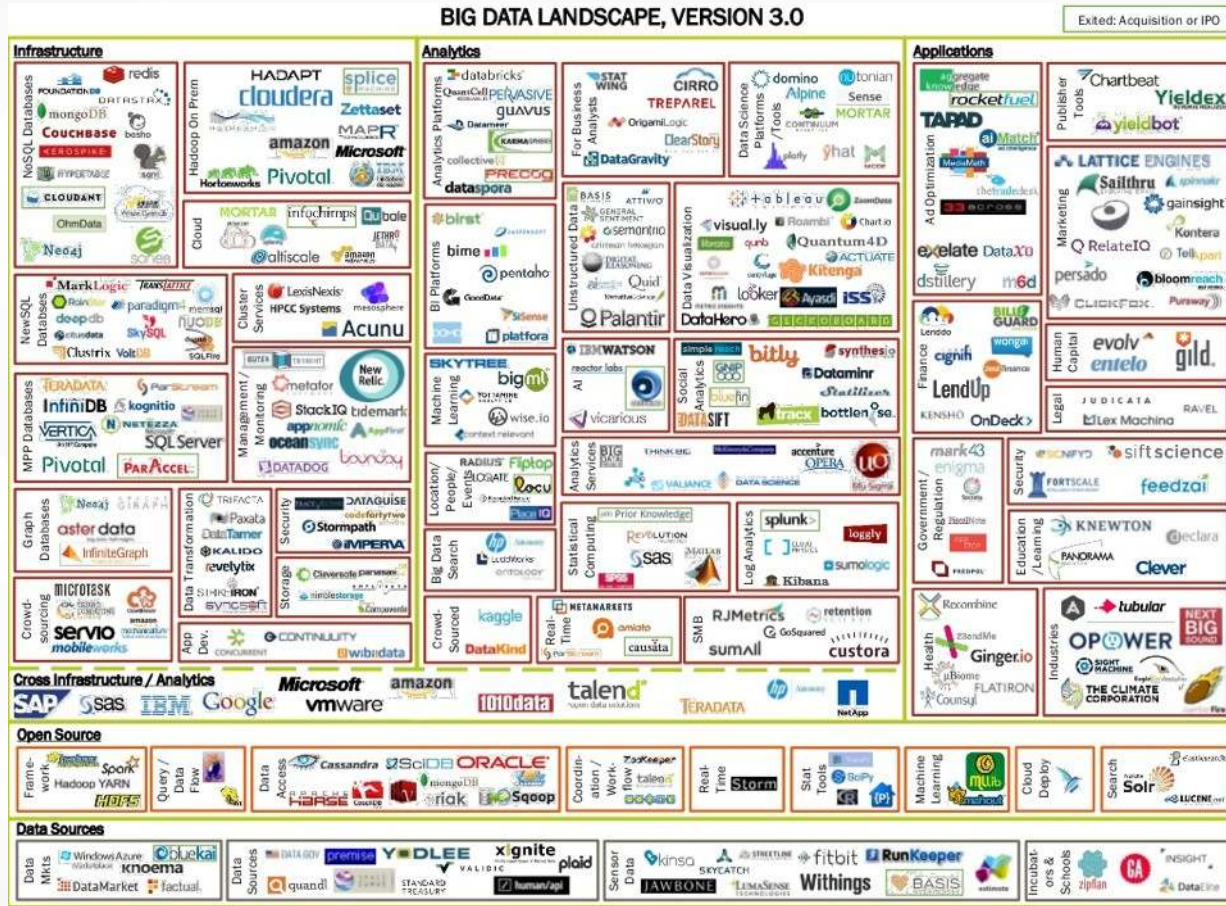
Julien Simon, Technical Evangelist, AWS

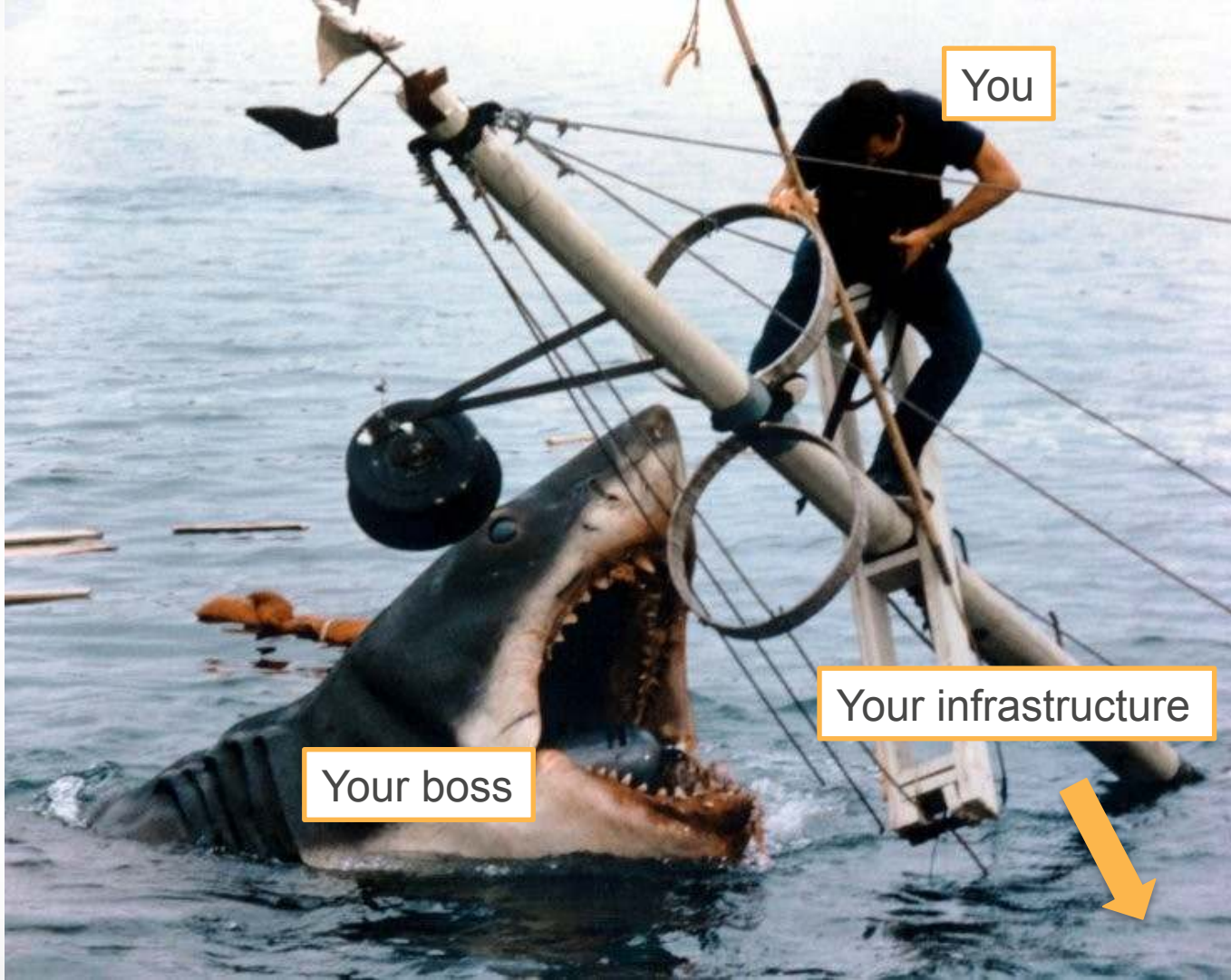
Lyon Data Science – 07/01/2016

julsimon@amazon.fr

@julsimon

Navigating the seven seas of Big Data





You

Your boss

Your infrastructure

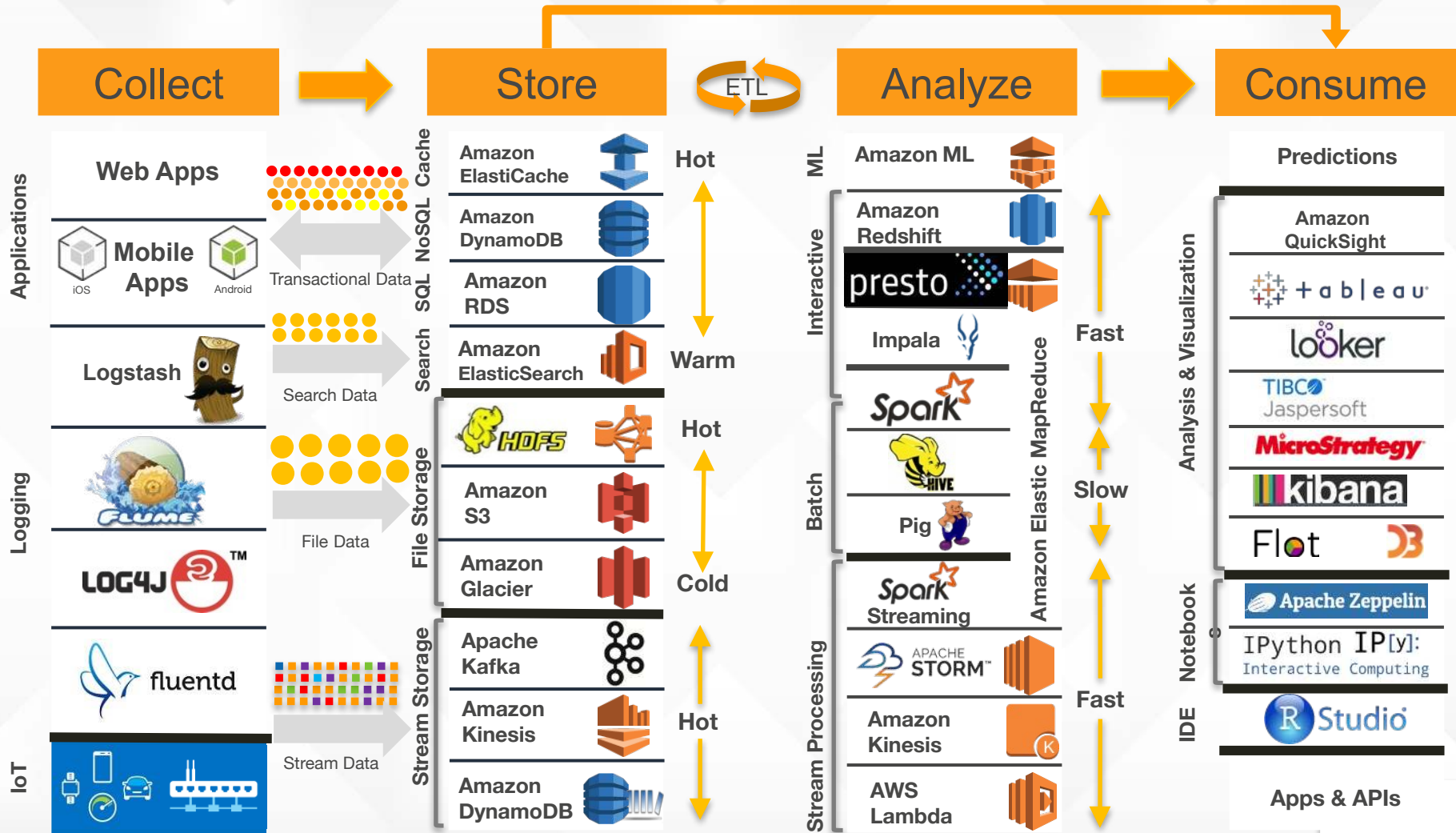


You need a better boat, not a bigger one

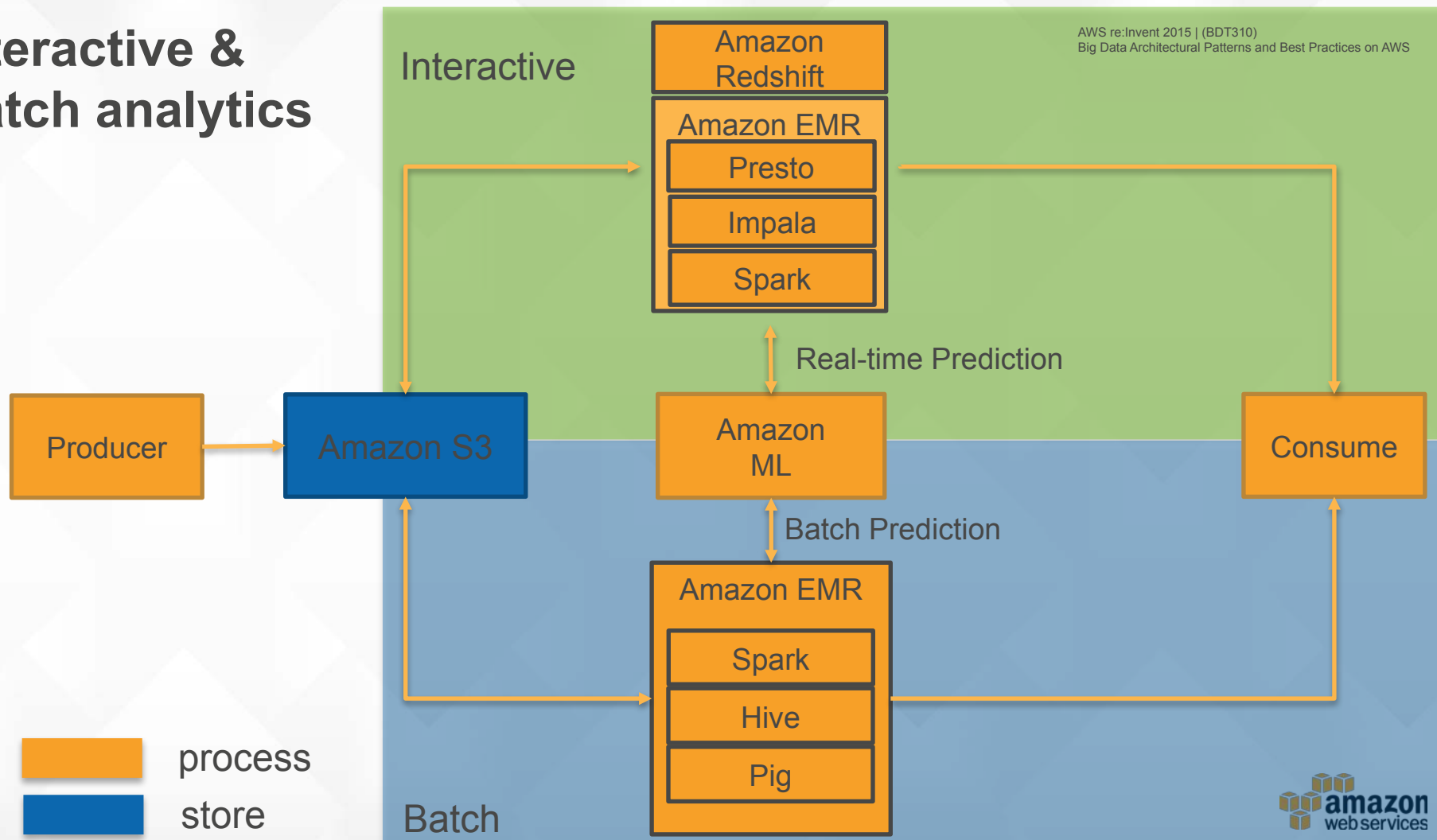
Let's face it, Big Data is great fun until:

- a) terabytes of data are lost forever
- b) tons of money are spent on non-scalable systems
- c) months are wasted on plumbing
- d) all of the above...

How about simple, cost-efficient, managed services that non-experts could use to build Big Data apps?

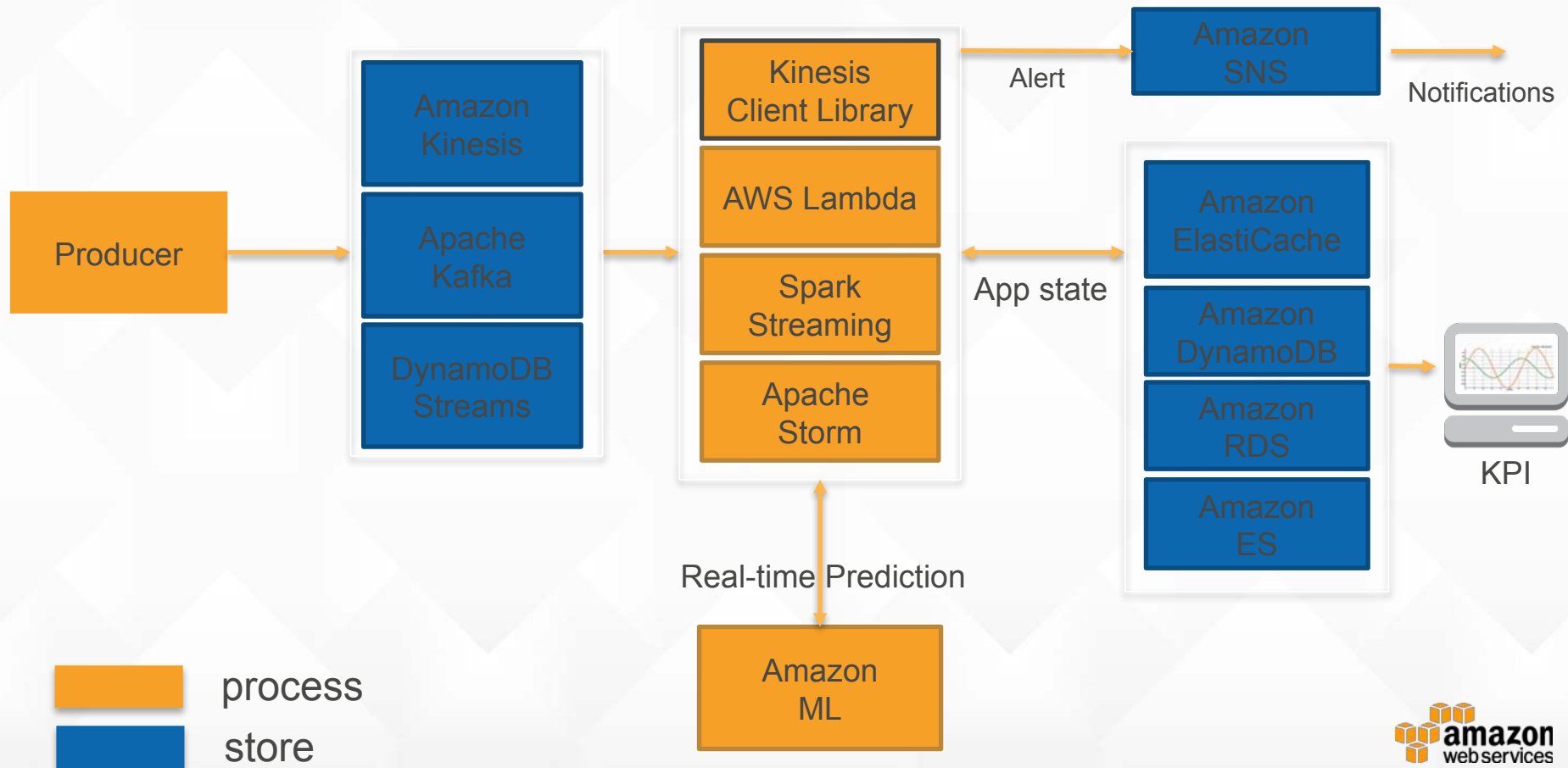


Interactive & Batch analytics



Real-time analytics

AWS re:Invent 2015 | (BDT310)
Big Data Architectural Patterns and Best Practices on AWS



Amazon Redshift

*an enterprise-level, petabyte scale, fully managed
data warehousing service*



Column-oriented database

Optimized for OLAP and BI workloads

Based on PostgreSQL 8.0.2

- SQL is all you need to know
- PostgreSQL ODBC and JDBC drivers are supported
- https://docs.aws.amazon.com/fr_fr/redshift/latest/dg/c_redshift-and-postgres-sql.html

Free tier: <https://aws.amazon.com/fr/redshift/free-trial/>

Available on-demand from \$0.25 / hour / node (us-east-1)

As low as \$0.094 / hour / node (us-east-1, 3-year RI)

Amazon Redshift architecture

Parallel processing

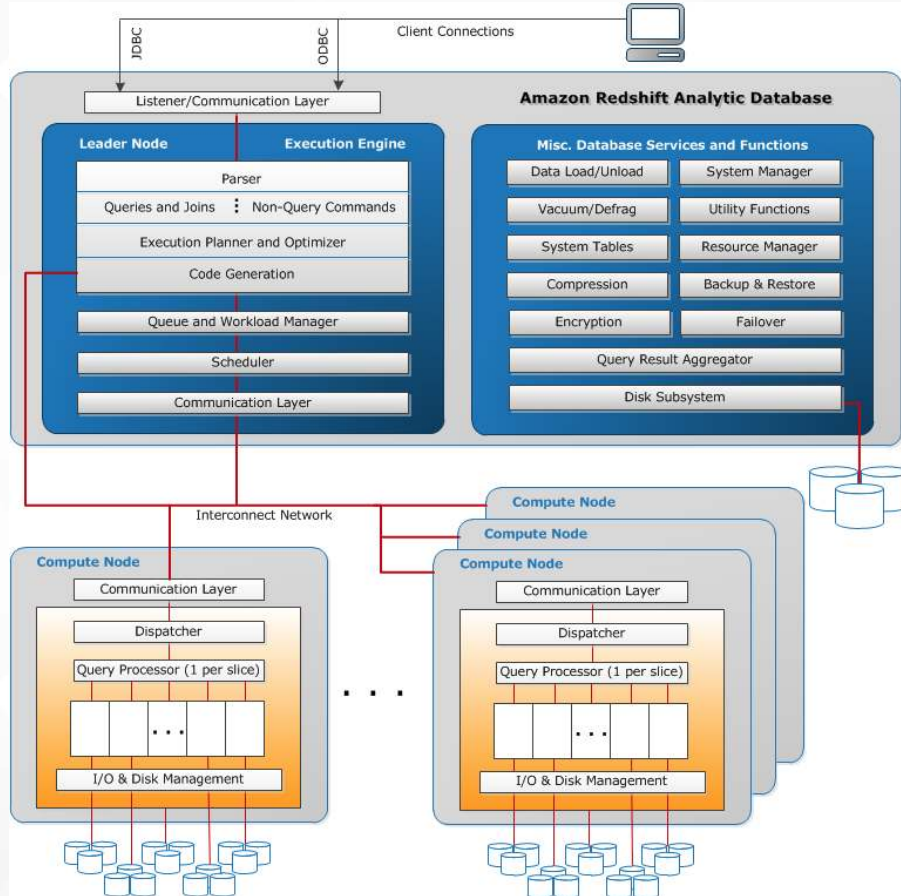
Columnar data storage

Data compression

Query optimization

Compiled code

Workload management



Instance types

Dense Storage Node Types

Node Size	vCPU	ECU	RAM (GiB)	Slices Per Node	Storage Per Node	Node Range	Total Capacity
ds1.xlarge	2	4.4	15	2	2 TB HDD	1-32	64 TB
ds1.8xlarge	16	35	120	16	16 TB HDD	2-128	2 PB
ds2.xlarge	4	13	31	2	2 TB HDD	1-32	64 TB
ds2.8xlarge	36	119	244	16	16 TB HDD	2-128	2 PB

Dense Compute Node Types

Node Size	vCPU	ECU	RAM (GiB)	Slices Per Node	Storage Per Node	Node Range	Total Capacity
dc1.large	2	7	15	2	160 GB SSD	1-32	5.12 TB
dc1.8xlarge	32	104	244	32	2.56 TB SSD	2-128	326 TB

Case study: Photobox



<http://www.lemagit.fr/etude/Photobox-consolide-et-analyse-ses-donnees-avec-AWS-RedShift>

Maxime Mezin, Data & Photo Science Director:

“L’entrepôt de données ne comportait que les données du site e-commerce liées aux ventes. Alors que nous avons la volonté d’intégrer des données du service clients et des données d’analyse (...) Nous avons atteint la limite du stockage de la base Oracle, et cela ne marchait pas très bien en termes de performances”

“Avec Redshift, la rapidité d’exécution des traitements a été multipliée par 10. Sans parler de la vitesse de chargement des données”

“On paie en fonction de la quantité de données que l’on va stocker. Chez Google, cela était plus compliqué”

- 2 Redshift clusters : 1 for historical data, 1 for real-time processing (SSD)
- TCO divided by 7 (90K€→13K€)

Case study: Financial Times

<https://aws.amazon.com/solutions/case-studies/financial-times/>



- BI analysis of customer usage, to decide which stories to cover
- Conventional data warehouse built using Microsoft technologies
- Scalability issues, impossible to perform real-time analytics → Amazon Redshift PoC
- Amazon Redshift performed so quickly that some analysts thought it was malfunctioning 😊

John O'Donovan, CTO:

“Some of the queries we’re running are 98 percent faster, and most things are running 90 percent faster (...) and the ability to try Redshift out before having to invest a significant amount of capital was a huge bonus.”

“Amazon Redshift is the single source of truth for our user data.”

“Being able to explore near-real-time data improves our decision making massively. We can make decisions based on what’s happening now rather than what happened three or four days ago.”



Amazon Redshift performance

No indexes, no partitioning, etc.

Distribution key

- How data is spread across nodes
- EVEN (default), ALL, KEY

Sort key

- How data is sorted inside of disk blocks
- Compound and interleaved keys are possible

Both are crucial to query performance!



Universal Pictures

DEMO #1

*Demo gods, I'm your humble servant,
please be good to me*

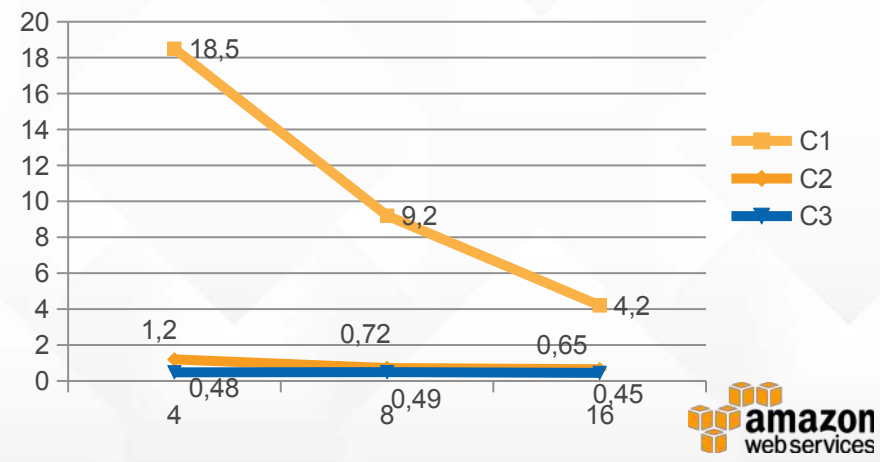
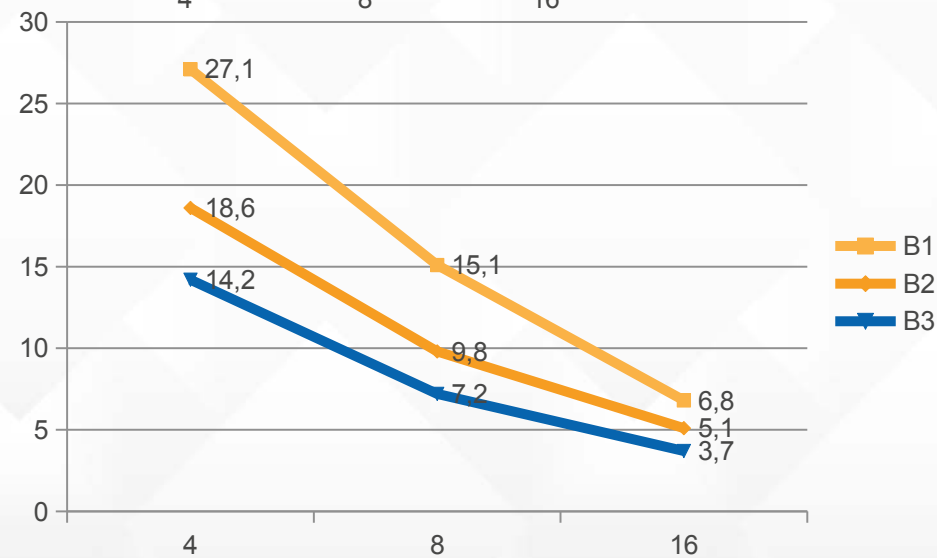
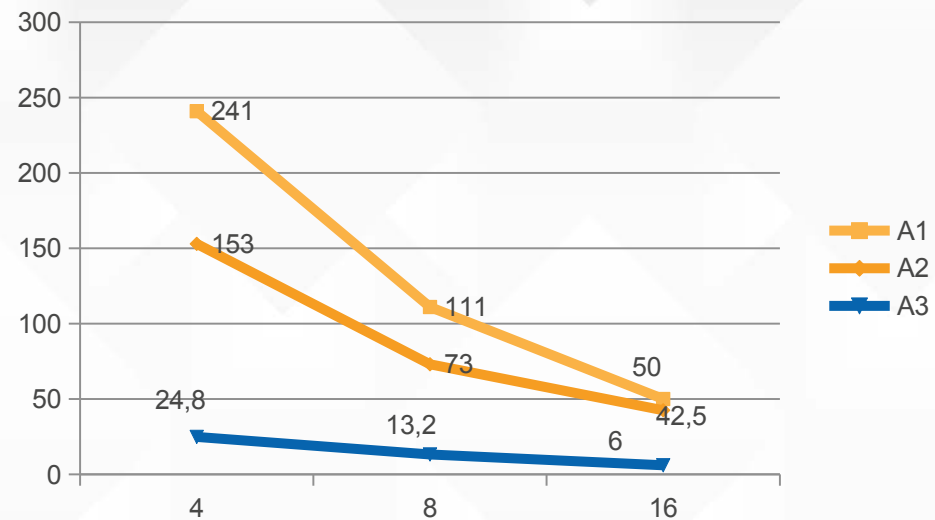
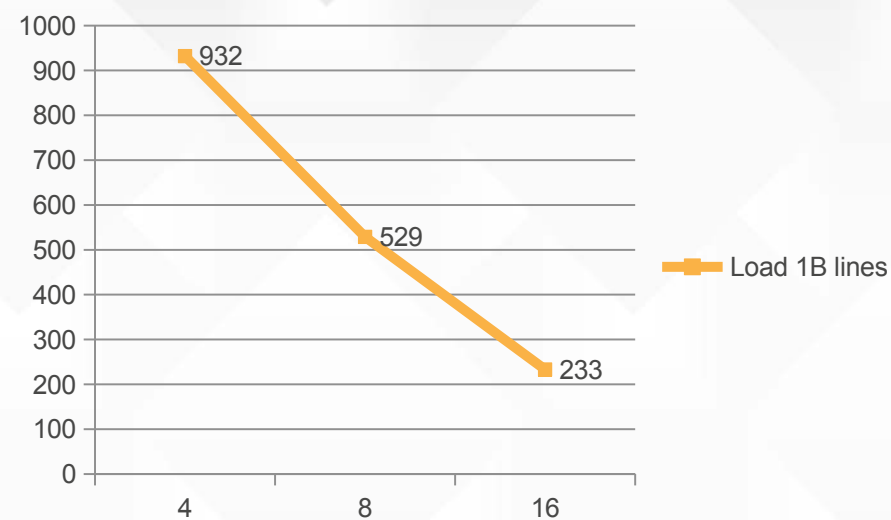
Create tables:
no key (1), sort key (2),
sort key + distribution key (3)

Load 1 billion lines (~45GB) from Amazon S3

Run queries (A, B, C) on a 4-node cluster

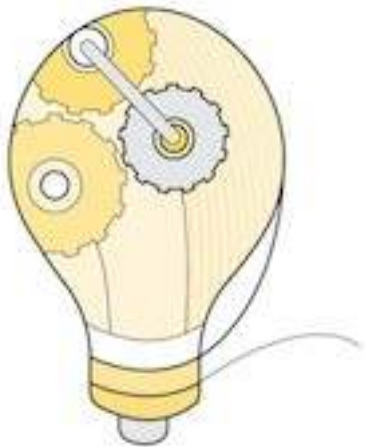
... while resizing the cluster to 16 nodes ☺





Amazon Machine Learning

*A managed service for building ML models
and generating predictions*



Integration with Amazon S3, Redshift and RDS
Data transformation, visualization and exploration
Model evaluation and interpretation tools
API for inspection and automation
API for batch and real-time predictions

\$0.42 / hour for analysis and model building (eu-west-1)
\$0.10 per 1000 batch predictions
\$0.0001 per real-time prediction

Amazon ML prediction algorithms

- Binary attributes → binary classification
- Categorical attributes → multi-class classification
- Numeric attributes → linear regression
- Code samples on <https://github.com/aws-labs/machine-learning-samples>

Case study: BuildFax

<https://aws.amazon.com/solutions/case-studies/buildfax/>

BuildFax:
On-Demand Property Condition.

More than
8,000
Cities &
Counties

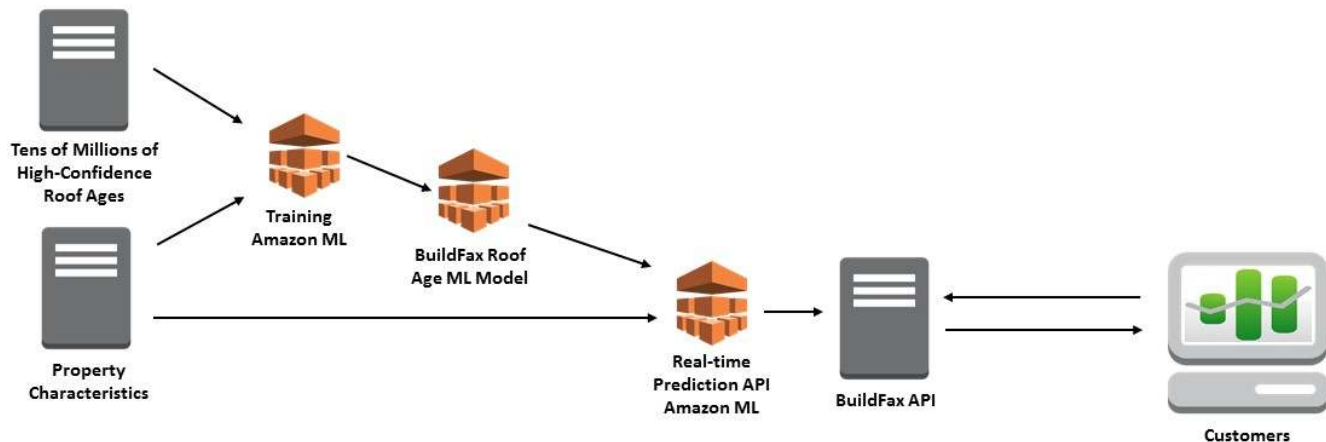
90+
MILLION
Properties

185+
MILLION
Building
Permits

10+
BILLION
Data Points

“Amazon Machine Learning democratizes the process of building predictive models. It's easy and fast to use, and has machine-learning best practices encapsulated in the product, which lets us deliver results significantly faster than in the past”

Joe Emison, Founder &
Chief Technology Officer



Other Amazon ML use cases



ICAO classifies and categorizes international aviation incidents daily



Plans to train ML models to predict fuel efficiency based on vehicle metadata, to push near-real-time data to customers



Securely sorts millions of mail pieces for clients every day for tremendous cost savings then brings it to the USPS for delivery

DEMO #2

*Demo gods, I know I'm pushing it,
but please don't let me down now*

Load data from Amazon S3 to Redshift and explore
Train and evaluate a regression model with Amazon ML
Perform batch prediction from data stored in Amazon S3

Load data from Amazon Redshift
Train and evaluate a regression model with Amazon ML
Create a real-time prediction API
Perform real-time prediction from Java app
(<https://raw.githubusercontent.com/juliensimon/aws/master/ML/MLSample.java>)



Thank you! Let's keep in touch 😊

Want to start a local AWS User Group? <https://aws.amazon.com/fr/usergroups/>

Many events and meetups in 2016 all across France

→ Please follow us at [@aws_actus](#) [@julsimon](#)

AWS Summit Paris : 31/05/2016 (free!) <https://aws.amazon.com/fr/paris16/>

13/01



14/01



BONUS SLIDES

Row storage vs columnar storage

SSN	Name	Age	Addr	City	St
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL

101259797 SMITH 88 899 FIRST ST JUNO AL	892375862 CHIN 37 16137 MAIN ST POMONA CA	318370701 HANDU 12 42 JUNE ST CHICAGO IL
-----------------------------------------	-------------------------------------------	------------------------------------------

Block 1	Block 2	Block 3
---------	---------	---------

SSN	Name	Age	Addr	City	St
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL

101259797 892375862 318370701	468248180 378568310 231346875 317346551 770336528 277332171 455124598 735885647 387586301
-----------------------------------	-----------------------------------------------------------------------------------------------------------

Block 1

Amazon Redshift & Machine Learning resources

Documentation

<https://aws.amazon.com/documentation/redshift/> <https://aws.amazon.com/documentation/machine-learning/>

Big Data videos from AWS re:Invent 2015

<https://blogs.aws.amazon.com/bigdata/post/Tx3D3UYOXB9XG6Z/Videos-now-available-for-AWS-re-Invent-2015-Big-Data-Analytics-sessions>

If you're going to watch only one: <https://www.youtube.com/watch?v=K7o5OIRLtvU>

More Amazon Redshift resources

Articles by Werner Vogel, CTO, Amazon.com

<http://www.allthingsdistributed.com/2012/11/amazon-redshift.html>

<http://www.allthingsdistributed.com/2013/02/amazon-redshift-resilience.html>

<http://www.allthingsdistributed.com/2013/05/amazon-redshift-designing-for-security.html>

Tuning Amazon Redshift

https://docs.aws.amazon.com/fr_fr/redshift/latest/dg/t_Sorting_data.html

https://docs.aws.amazon.com/fr_fr/redshift/latest/dg/t_Distributing_data.html

<http://blogs.aws.amazon.com/bigdata/post/Tx31034QG0G3ED1/Top-10-Performance-Tuning-Techniques-for-Amazon-Redshift>

Creating and deleting an Amazon Redshift cluster

```
$ aws redshift create-cluster --cluster-identifier CLUSTER_NAME  
--node-type dc1.large --number-of-nodes 4  
--db-name DATABASE_NAME  
--master-username USER_NAME  
--master-user-password USER_PASSWORD  
--publicly-accessible          ← probably not what you want!
```

```
$ aws redshift delete-cluster --cluster-identifier CLUSTER_NAME  
--skip-final-cluster-snapshot  ← data will be lost!
```


Connecting to Amazon Redshift with *psql*

```
$ psql -h xxx.redshift.amazonaws.com -p 5439  
-d DB_NAME -U USER_NAME
```

Force SSL:

```
$ psql -h xxx.redshift.amazonaws.com -p 5439  
-U USER_NAME "dbname=DB_NAME sslmode=require"
```

Loading data from Amazon S3 to Amazon Redshift

COPY command example

```
$ copy TABLE_NAME  
from 's3://BUCKET_NAME/FOLDER_NAME'  
region 'eu-west-1'  
credentials 'aws_access_key_id=MY_ACCESS_KEY;  
aws_secret_access_key=MY_SECRET_KEY'  
delimiter ',' bzip2 maxerror 1000;
```

View last 10 load errors

```
select * from stl_load_errors order by starttime desc limit 10;
```

Resizing an Amazon Redshift cluster

```
$ aws redshift modify-cluster  
--cluster-identifier mycluster  
--number-of-nodes 8
```

Listing Amazon ML models

```
$ aws machinelearning describe-ml-models  
--query "Results[*].{Name:Name, Id:MLModelId,  
Type:MLModelType}"
```