Get Your Guide

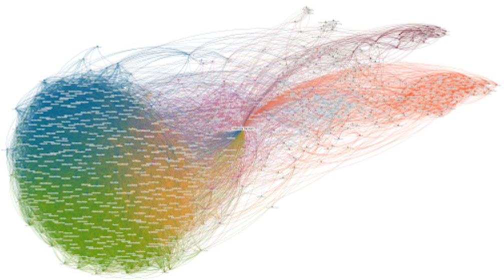
Mathieu Bastian
June 12th 2017 - Berlin Buzzwords

365 days of Apache Spark™

BERLIN BUZZWORDS 2017 JUNE 11-13

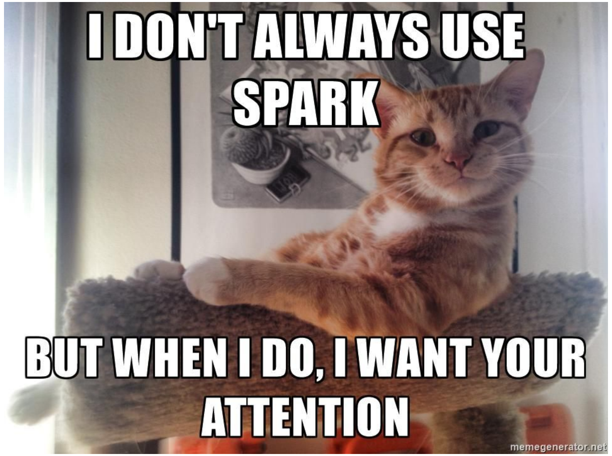
About me

- Data scientist and data engineer, all data matters!
- Paris ➟ San Francisco ➟ Berlin
- Led data products team at LinkedIn
- Co-founder of Gephi open-source software
- Head of Data at GetYourGuide



About this mission

- GetYourGuide is the leading global marketplace for tours and activities
  - Scale-up of 350+ employees, based in Berlin
- Types of datasets
  - User behavior (e.g. events)
  - Transactional data (e.g. bookings, payments)
  - Performance marketing (e.g. keywords, impressions)
  - Images, reviews, geolocations etc.
- Started at GetYourGuide in Feb 2016
  - Data mostly organized around single Data Warehouse
  - Your mission: Build a new data platform
  - Mission accepted! *Can I use Spark?*



http://www.fitnesslibre.com/wp-content/uploads/2013/12/Games2012_WomenObstacleCourse.jpeg

A unified data platform

The end of the continental divide

Get Your Guide

Two fundamental goals

Data ➡ decisions

- ❑ Metrics, reports and dashboards
- ❑ Deep-dive insights (exploratory)
- ❑ Data visualization

Building data products

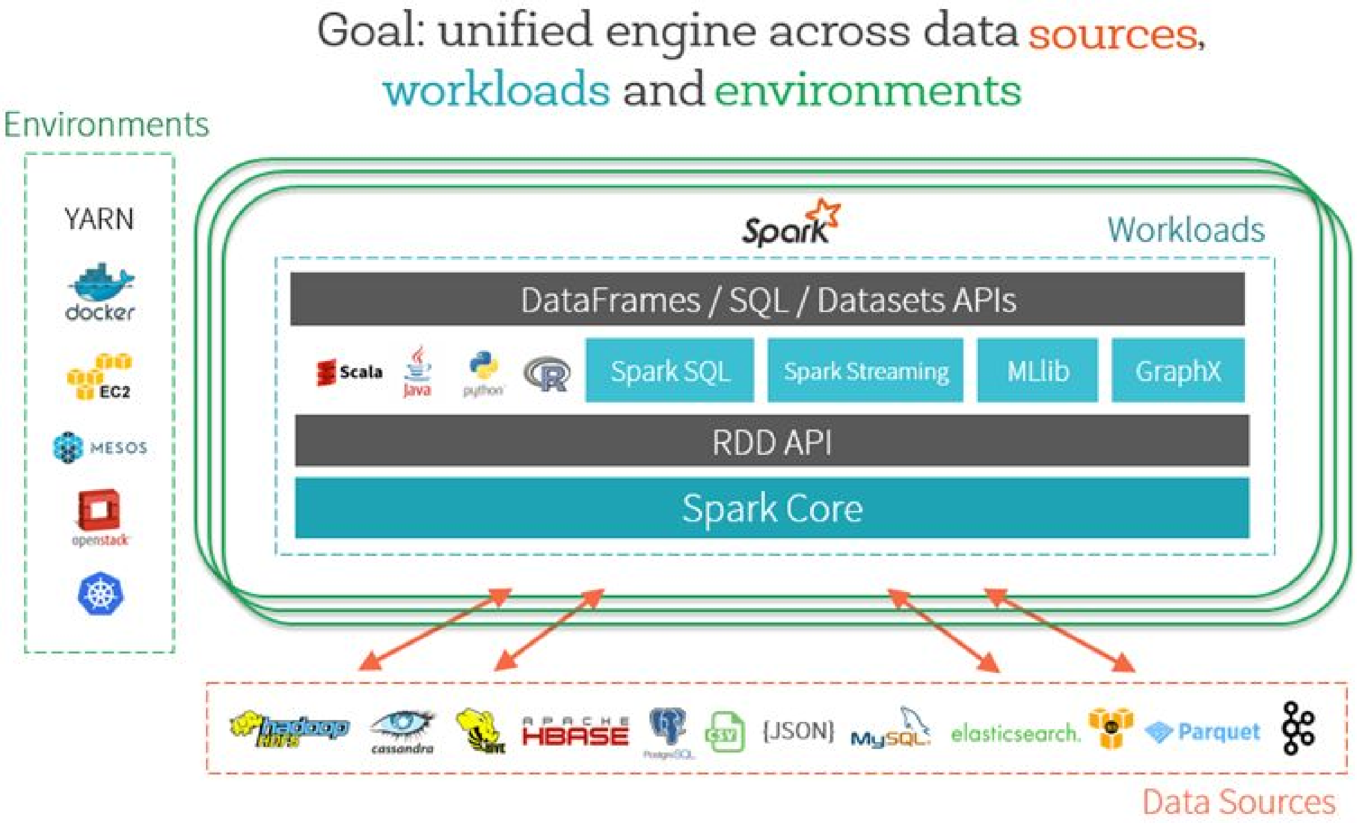
- ❑ Algorithms and Machine Learning
- ❑ Many different sources and formats
- ❑ Fully automated and reliable

Also those goals...

- ❏ Keep architecture future-proof
- ❏ Scale gracefully to large datasets and more complex use-cases
- ❏ Fast to setup (we're a startup!)
- ❏ Build infrastructure incrementally, while still delivering

Get Your Guide

We're aligned!



Source: Spark documentation

Get Your Guide

Pick the right tool for the job

- What do others say?

Spark

Audience

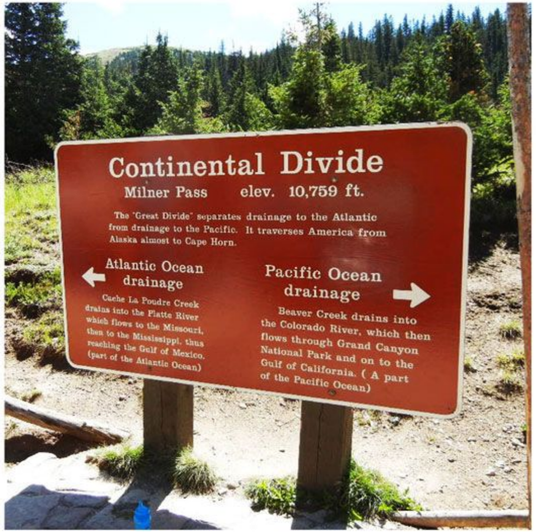
	Apache Hive	Apache Impala (incubating)	Apache Spark SQL
Audience	ETL Developers	Business Analysts	Data Engineers & Data Scientists
Strengths	• Built for very long-running ETL, data preparation, or batch processing • Supports custom file formats • Handles massive ETL sorts with joins	• Scales to high-concurrency • Supports high-performance interactive SQL • Compatible with BI tools & skills • Hadoop integration & usability	• Easily embed SQL into Java, Scala, or Python applications • Simple language for common operations • Seamlessly mix SQL and Spark code within a single application
New Features	• Hive in the cloud (S3) • Hive-on-Spark beta • Governance & Lineage	• Nested data types • Column-level security • Integration with Kudu (beta)	• Support for Spark SQL & DataFrames • Hive integration • Automatic performance optimizations

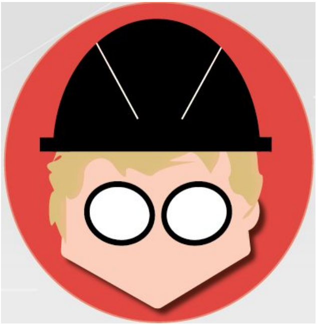
Source: https://www.slideshare.net/cloudera/hive-impala-and-spark-oh-my-sqlonhadoop-in-cloudera-55

Pick the right tool for the job

Data Scientist

Data Analyst



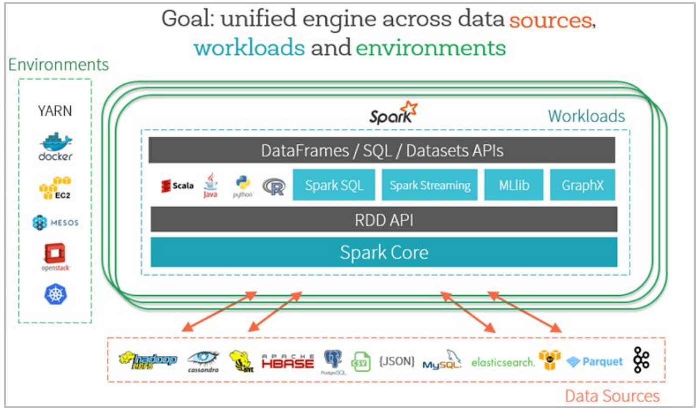


Data Engineer


I WANT IT
ALL
AND
I WANT IT
NOW

The 3 reasons why it works

- Interactive querying
  - SQL (Ansi SQL)
  - Small task ~= Small runtime
  - Progress vs Spinner
- Standardized, rich API
  - From prototype to production
  - Standard machine learning library (distributed)
- Easy integration
  - Interoperability with other tools
  - Data sources and connectors
  - Streaming capabilities

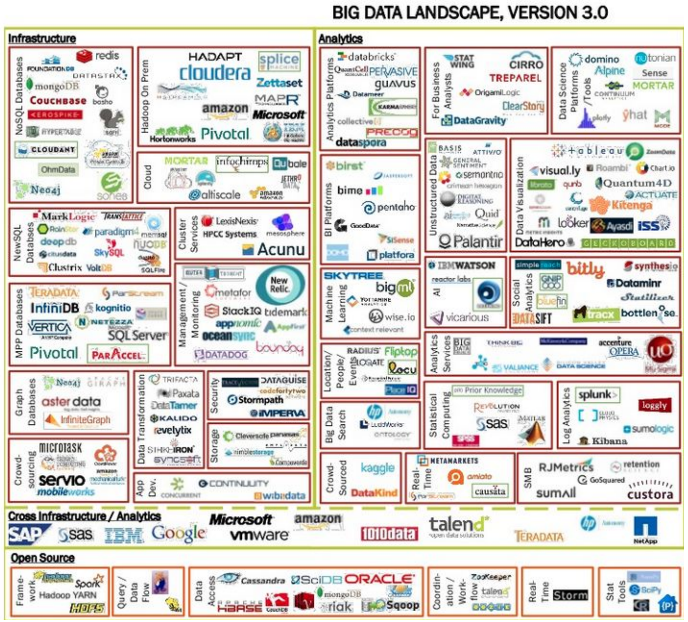


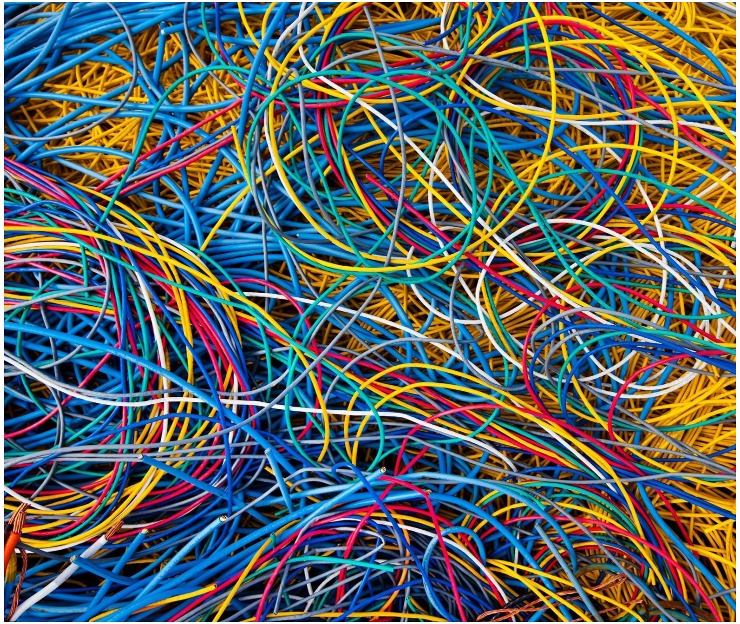
A nimble data platform

Simplicity and flexibility

Get Your Guide

When I think about big data platforms...





What I really want

Download on the App Store

GET IT ON Google play

Get Your Guide

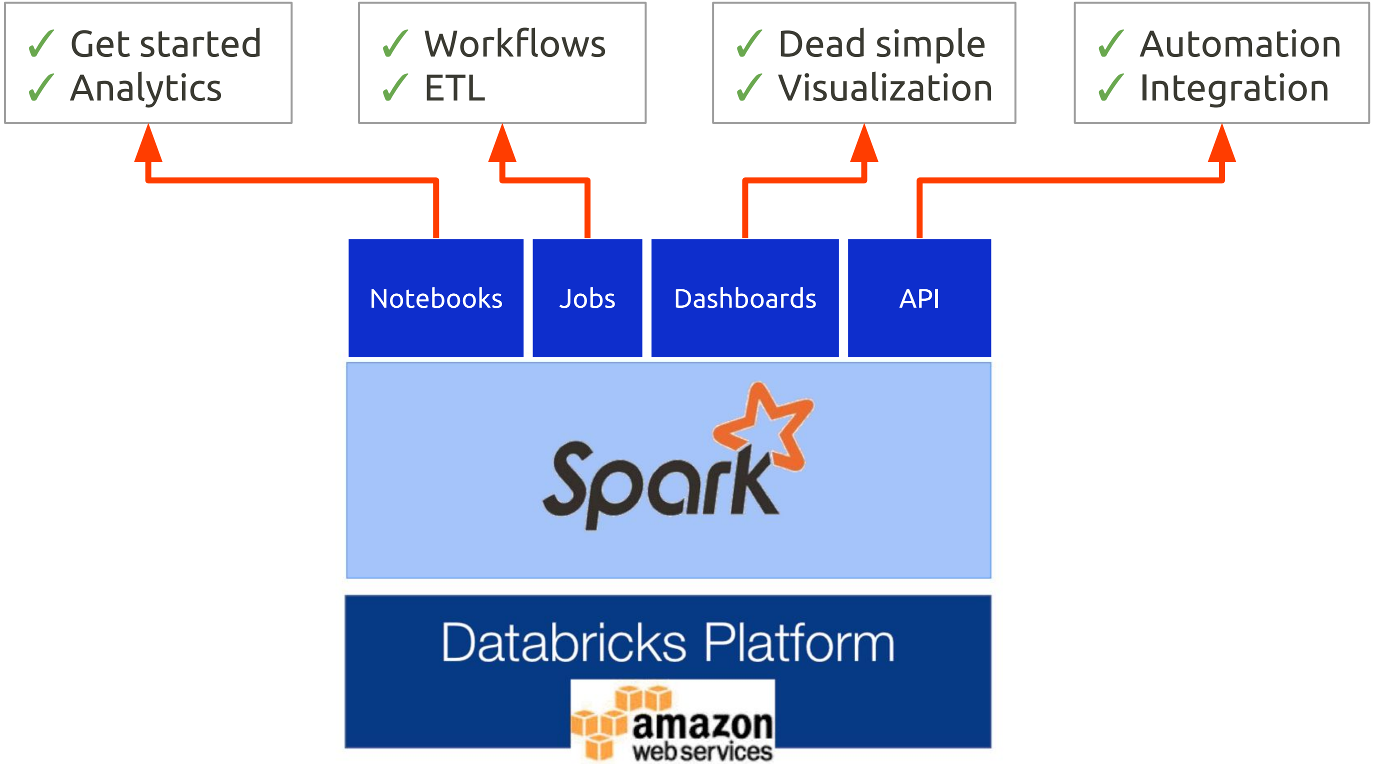
New data platform

- MVP mindset
  - Easy and quick to setup
  - Iterative improvements

- Databricks in the cloud
  - Cloud provider for Apache Spark
  - Founded by creators of Spark
  - Sits on top of AWS
  - Multiple clusters management

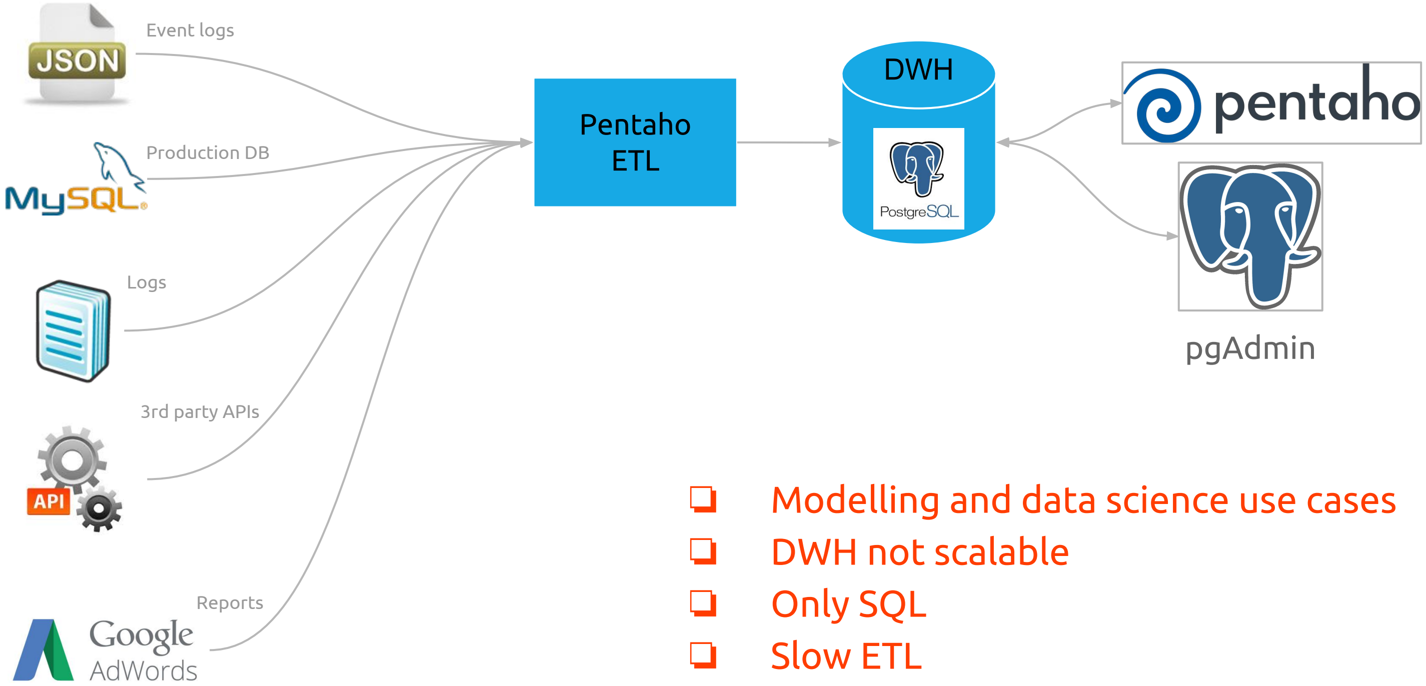
databricks

Databricks


✓ Get started
✓ Analytics
✓ Workflows
✓ ETL
✓ Dead simple
✓ Visualization
✓ Automation
✓ Integration
Notebooks
Jobs
Dashboards
API
Spark
Databricks Platform
amazon
web services

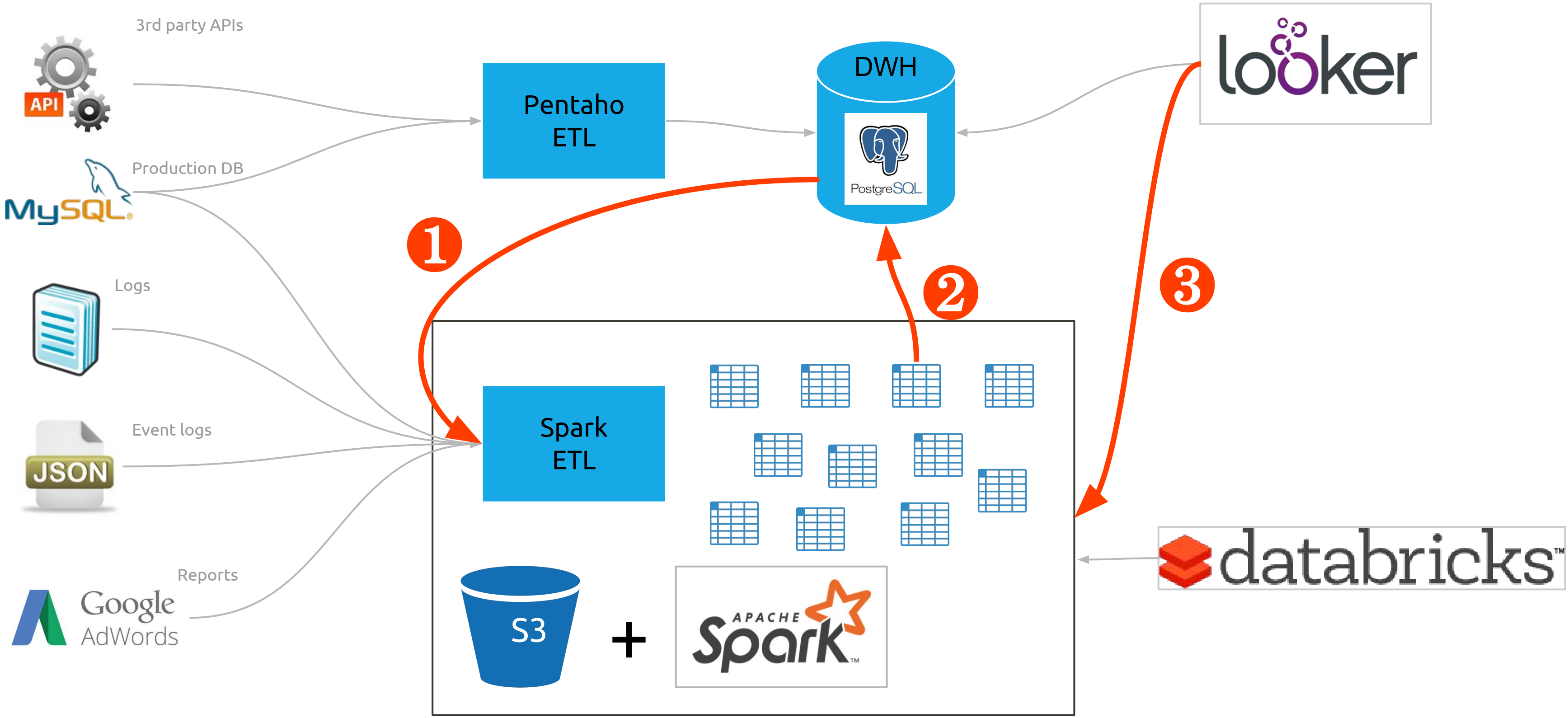
Get Your Guide

A while ago...



- ❏ Modelling and data science use cases
- ❏ DWH not scalable
- ❏ Only SQL
- ❏ Slow ETL

From good to great!



365 days of Spark use-cases

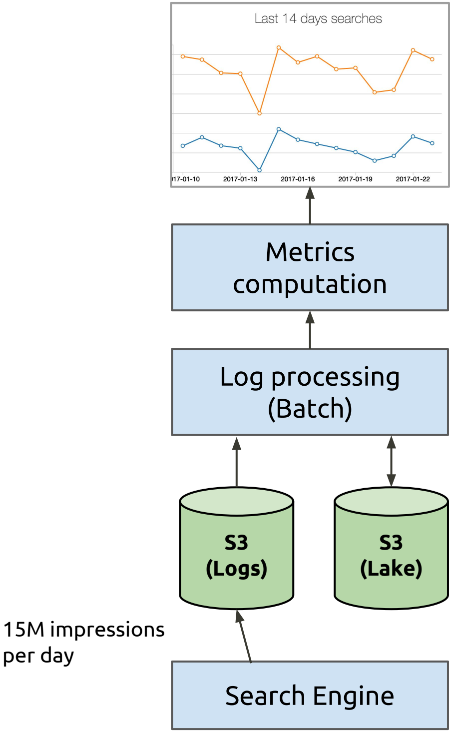
Search Dashboarding

Performance Management

Offloading Aggregations

365 days of Spark use-cases

Search Dashboarding



Get Your Guide

365 days of Spark use-cases

Performance Management

	Legacy	Now
Setup	On-premise server	Spark 2.1
Frequency	Once a week	Twice a day
Data size	1x	100x
Storage	PostgreSQL	Parquet

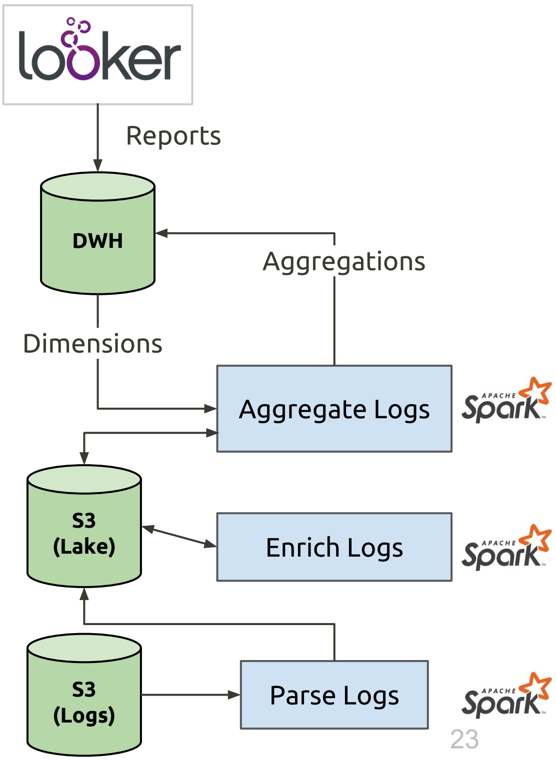
APACHE Spark

+

Google AdWords

365 days of Spark use-cases

Offloading Aggregations



Get Your Guide

The storage question

- Anticipated growth pains with storage
  - Cost out of control
  - Lack of structure in formats and schemas (e.g. CSVs)
  - Redundant data for each use-case
  - Impact on Spark performance

Data Lake!

Get Your Guide

- Data Lake philosophy
  - Save raw data now to analyze later
  - Centralisation brings efficiency
  - Access for everyone

- Spark ↔ Data Lake
  - Parquet format
    - Performance + Long-term storage
    - Interoperability (future proof)
  - Tables === Files

Avoid data clutter

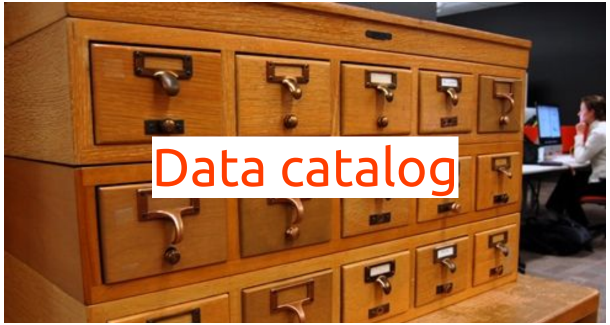
- ❑ Schemas!
- ❑ Data classification
- ❑ Discovery and search

Catalog & Metadata	Flexible Access	Security & compliance

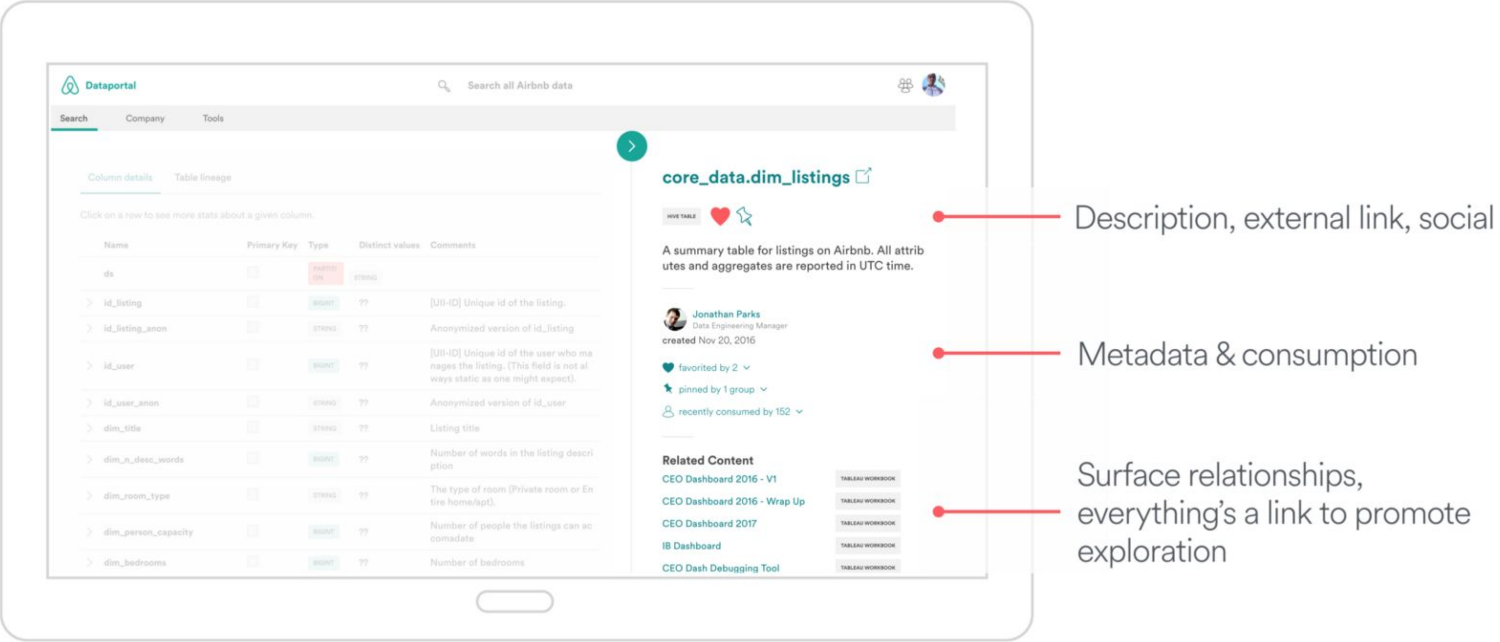
Parquet

Storage

VS



Data discovery



AirBnB's dataportal

Source: https://medium.com/airbnb-engineering/democratizing-data-at-airbnb-852d76c51770

Future proof platform

“ Premature optimization is the root of all evil ” - *Donald Knuth*

- Solutions
  - Rely on a unique open-source, standard technology
  - Spark API, interoperable formats

Onboarding strategy

Let's talk about people!

Get Your Guide

What are our goals again?

- Goals
  - People are **empowered** to make data-driven decisions
  - People can find **clean data** to work with
  - People can **innovate** rapidly in building data products

- Challenges
  - Friction in accessing and analyzing data
  - Eliminate the crutch, be truly self-service
  - The vast majority or data users unfamiliar with Spark
  - Anticipate data science needs

Lessons learnt

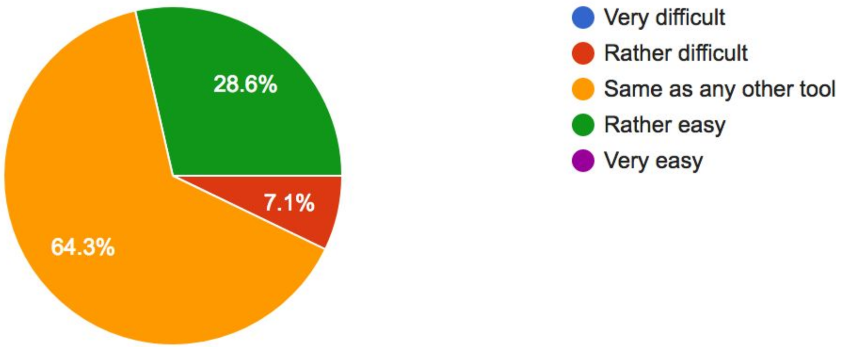
- Bring data
  - Data Warehouse tables early
  - Make it super easy and fast to add new tables (and avoid tickets)
- Early conventions
  - Parquet
  - Path structure
- Training and examples
  - Educational Scala notebooks

- Data source changes
  - Events restructure
  - Long-term history needed for analysis and insights
- Data quality
  - Trust is hard
  - Numbers won't match
- Mixing Python and Scala
  - Code duplication and libraries

Learning

Compared to other data tools you have worked with before, how difficult is it to learn Spark?

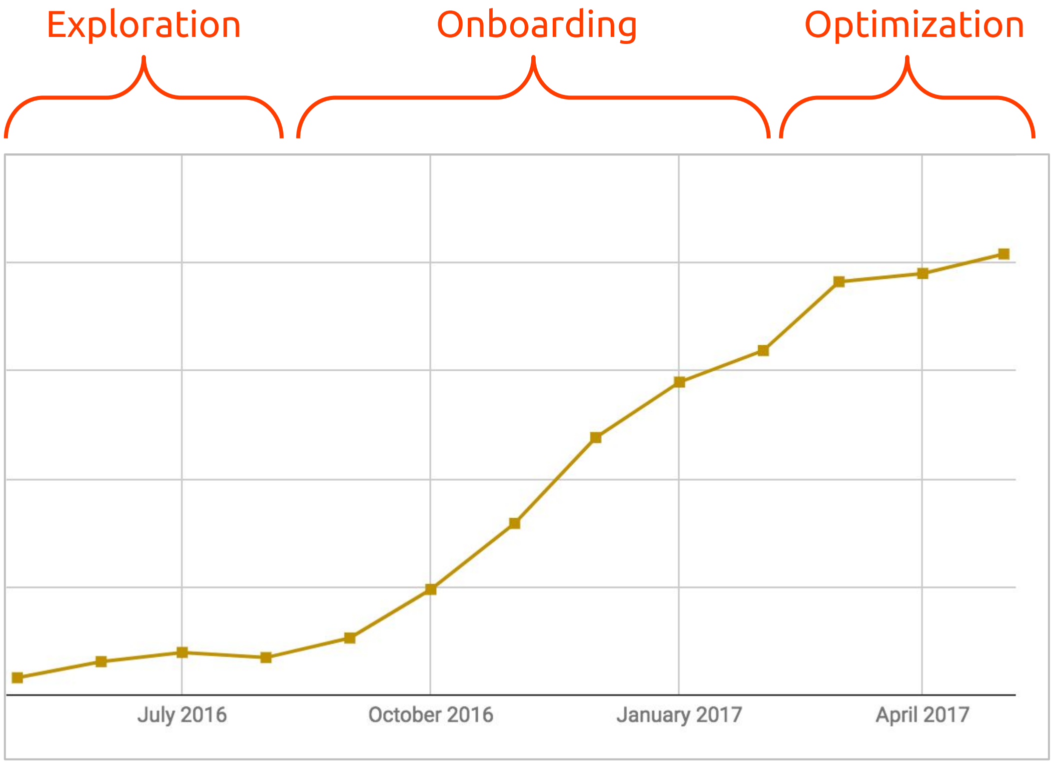
14 responses



The hard work

Post onboarding

Post onboarding



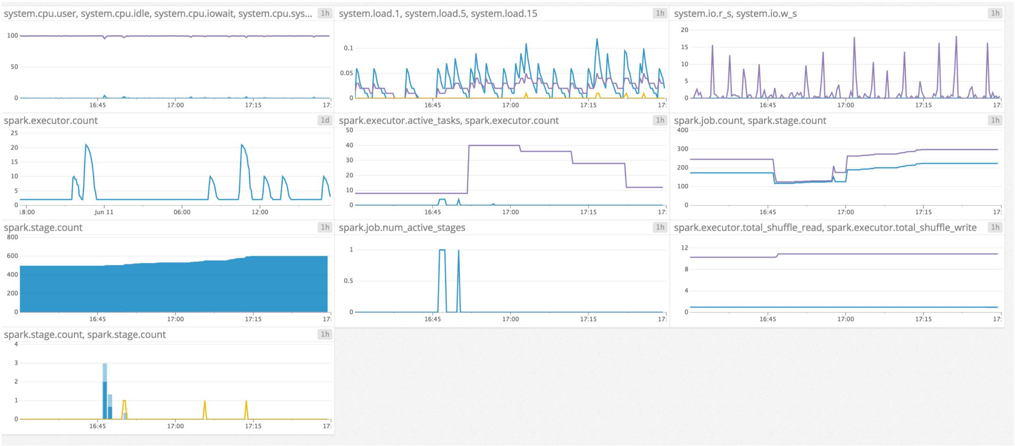
Monthly
Spark Usage

The main challenges we faced

- ❏ Getting real with cluster administration
- ❏ Deeper understanding of performance factors
- ❏ Understanding root causes
- ❏ Organizing data dependencies
- ❏ Ensuring data quality and standardization

Cluster administration

- ❏ Common issues
  - ❏ Driver crashing
  - ❏ Lost executors
- ❏ Built connector to DataDog
  - ❏ Hard to estimate capacity
  - ❏ Navigate into confusing metrics
- ❏ Cluster start/stop
  - ❏ Autoscaling



DataDog metrics cluster monitoring

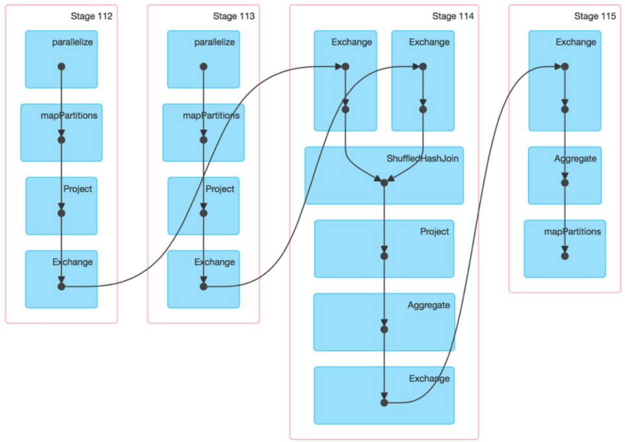
Debugging inquiries

- Logs are hard to read/process
- SparkUI is useless for the most part
- Can't easily detect problems (e.g. memory problems)

Details for Job 8

Status: SUCCEEDED
Completed Stages: 4

▸ Event Timeline
▾ DAG Visualization



Spark Stages

Total Duration: 1.5 m
Scheduling Mode: FIFO
Active Stages: 2
Completed Stages: 3
Failed Stages: 0

Active Stages (2)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Shuffle Read	Shuffle Write
3	flatMap at ALS.scala:326	2014/02/19 05:39:31	17.0 s	97/120		2.3 GB
6	map at ALS.scala:147	2014/02/19 05:39:30	18.0 s	77/183		33.7 MB

Completed Stages (3)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Shuffle Read	Shuffle Write
5	mapPartitionsWithIndex at ALS.scala:164	2014/02/19 05:39:30	1.0 s	120/120		662.9 MB
0	reduceByKeyLocally at ALS.scala:217	2014/02/19 05:39:20	10.0 s	120/120	63.2 MB	
1	map at ALS.scala:146	2014/02/19 05:38:50	29.5 s	183/183		79.0 MB

Failed Stages (0)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Shuffle Read	Shuffle Write

Log processing on Spark

- Legacy solution based on Pentaho Data Integration
  - Configuration vs Code
  - Scalability challenges
- Migrate to Spark
  - Data quality challenges

Log parsing workflow in Pentaho

Get Your Guide

Log processing on Spark

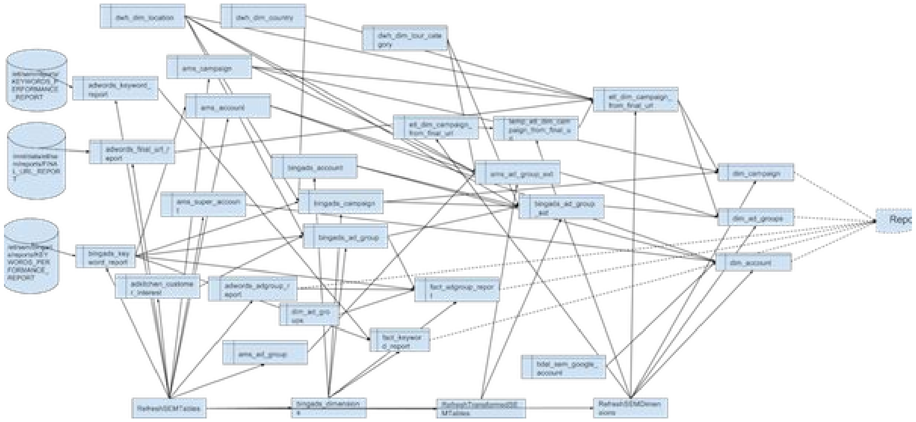
client_ip	String	length<=50 255.255.255.255 (15 characters) ip6 2001:0db8:85a3:0000:0000:8a2e:0370:7334 (39 characters)	N	ERR	<field> is null or out of range
country_iso_code	String	lentgh=2	Y	WAR	<field> is null or out of range
currency	String	lentgh=3	Y	WAR	<field> is null or out of range
date_time	Timestamp	YYYY-MM-DDTHH24:MI:SS	N	ERR	<field> is null or out of range
partner_id	String	length<=32	Y	ERR	<field> is out of range
partner_src	Integer	IN (1,2,3,4,5,6)	Y	ERR	<field> is out of range
partner_cmp	String	length<=100	Y	ERR	<field> is out of range
platform	String	IN ('mobile','desktop') OR NULL	Y	ERR	<field> is out of range
request_url	String		N	ERR	<field> is null
referrer_url	String		Y	NONE	

- Scala library
  - Unit and integration testing
  - Easier to benchmark
- Validation, testing and errors
  - Incremental severity (warning, errors)
  - Edge cases
  - Track errors

error
[{"column_name":"country_iso_code","error_type":"warning","error_message":"is_null"},{"column_name":"currency","error_type":"warning","error_message":"is_null"}]
[{"column_name":"country_iso_code","error_type":"warning","error_message":"is_null"},{"column_name":"currency","error_type":"warning","error_message":"is_null"}, {"column_name":"session_id","error_type":"error","error_message":"is_null"},{"column_name":"visitor_id","error_type":"error","error_message":"is_null"}, {"column_name":"locale_code","error_type":"warning","error_message":"is_null"}]

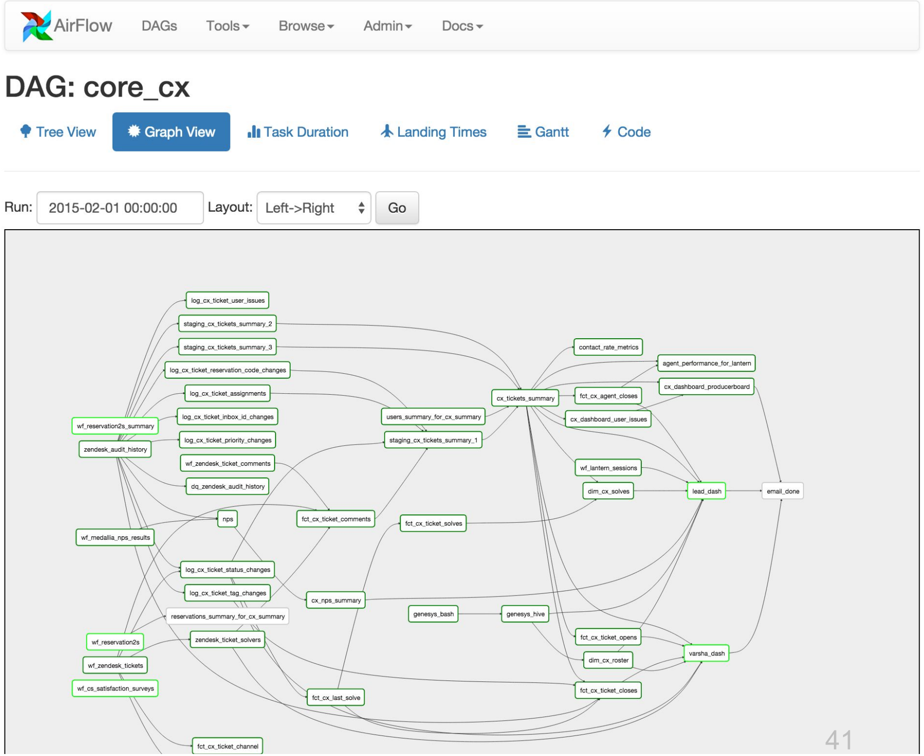
Workflow orchestration

- Data lineage
  - Recovery and SLAs
  - Data dependencies
- From 10 to 100 jobs
  - Self-service, undeclared consumers
  - Documentation and onboarding
  - Cluster utilization



Workflow orchestration

- Apache Airflow
  - Map out data dependencies
  - Flexible configuration
  - Backfilling and data management
  - Operators
    - Databricks
    - PostreSQL
    - …

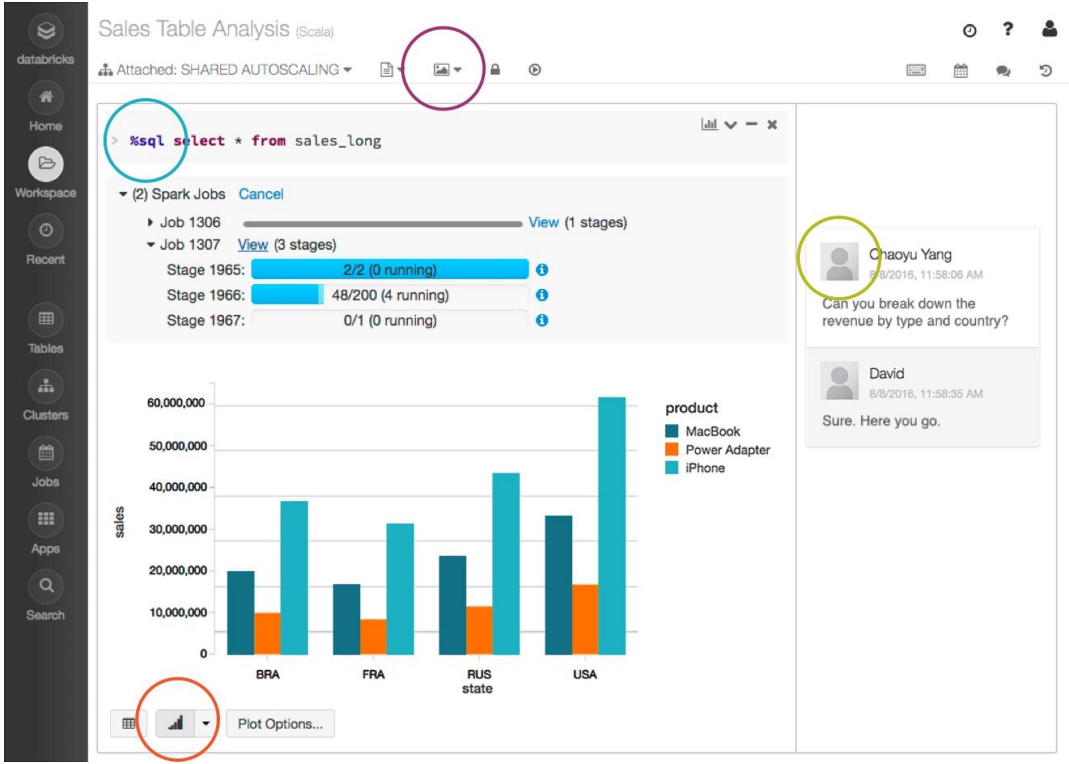


Notebooks

The  and 

Notebooks

Get Your Guide



- Contained collection of queries or code snippets

Scala SQL R python™

- Data presentation and visualization

Tables

Visualization

Get Your Guide

The obvious advantages

- **Iterative development**
  Chances you'll get your code or query right at the first try is close from zero
- **Exploratory data analysis**
  Use simple visualizations (e.g. histogram, line chart) to ask questions to the data
- **Visible and collaborative**
  Code and analysis aren't buried into Git repositories but easy to discover and review
- **Easy to get started and learn**
  Online, safe environment to get started with Spark concepts and syntax
- **Also open-source**
  Apache Zeppelin also easy to get started with

Get Your Guide

But you can also do...

- Run a notebook with parameters as part of a workflow

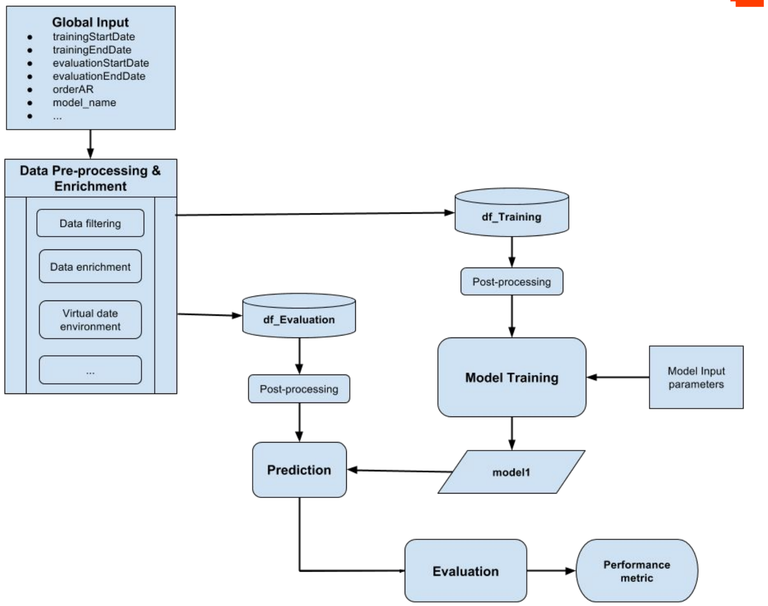
```
> var returnvalue = dbutils.notebook.run("./notebook2", 60, Map("data" -> "records"));
```

Notebook job #6120

- Run notebooks as part of other notebooks
- Develop utilities and libraries in notebooks
- Synchronize your notebook on Git repositories
- Use Databricks' notebook API
- Send execution logs to Sentry and by Email
- Use multithreading to run notebooks in parallel

Get Your Guide

Production workflows the right way



- Notebooks are too limited to scale production workflows
  - Ability to design unit and integration tests
  - Proper revision history and code review
  - Global configuration
  - Separate development from production environment
  - Multi-module projects with dependencies
  - Complete control on error handling and logging

Looking back...

Never been this easy to build large-scale production workflows!

- Compared to Hadoop
  - Large overhead and complexities in testing locally
  - No proper investment in unit-testing (MRUnit)
  - Mix multiple languages (not only Java)
- Compared to Pig
  - Built around simplistic data structures (Text vs Avro)
  - Cumbersome mocking and testing

Wrapping-up

Thank you for your attention!

mbastian@getyourguide.com

We're hiring!

https://careers.getyourguide.com/