

# *War stories with Apache Spark*

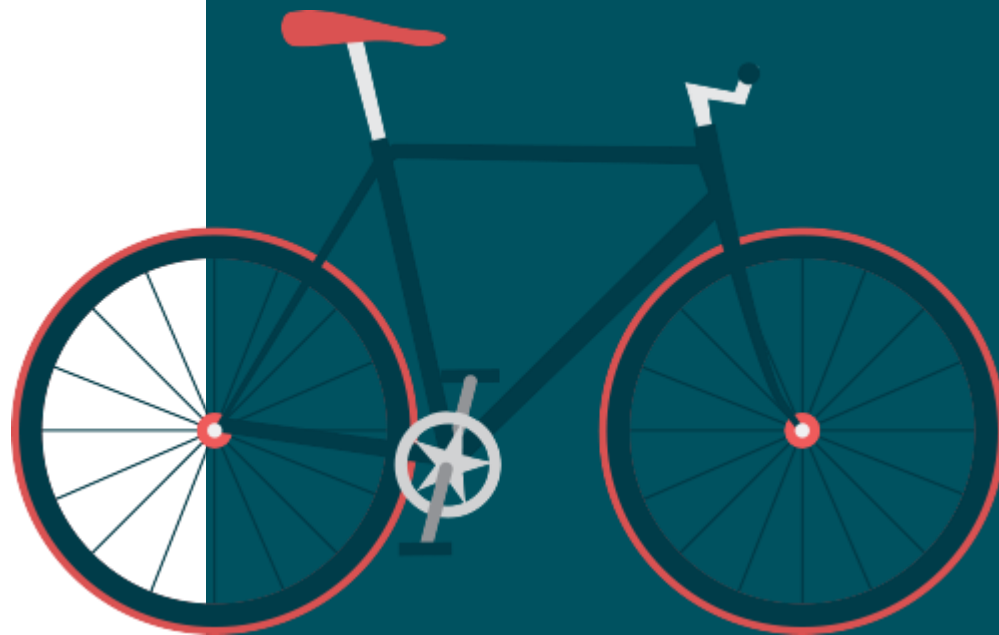
Mate Gulyas



# GULYÁS MÁTÉ

CTO & Co-Founder

 [@gulyasm](https://twitter.com/gulyasm)



2015Q1 Full automation - account mgmt nélküli orv. indítás

BIG DATA

GUI

SITE

- automata "stack" gen.
- stack planning
- code auto-gen.
- self-test (test/runner)
- auto scaling infra (?)

- ↔ user mgmt ↔
- ↔ user/pkg def. ↔

registrações felület

↓

2014 Q4 STAB

- tracking gyerekbetegsége
- pl: www. / ?
- lent aszimmetrikus
- ktsy csök.
- CPA / CPL fraud tracking

- DB strukt.
- GUI gyerekb./redesign
- chart usability megújítás
- funkciók (+)
- több szintű chartok
- TODO brief + use case

- M. O.
- ↳ use case (bot vs hum.)
- ↳ case study
- ↳ referenciák
- Blog
- Social int.

tracur  
on  
nemzetköz.

EVLI

FIBAN®  
FINNISH  
BUSINESS ANGELS  
NETWORK

SLUSH

SLUSH 100 PITCHING COMPETITION 2014

Date: 19.11.2014

Pay to

Euros

Five hundred thousand euros

€ 500,000.00

SLUSH 100 PITCHING COMPETITION WINNER

ENBRITE.LY



# DEMONSTRATION REPORT

Data as of March 7th at 4:12pm



- HOME
- MY DASHBOARDS
- Clients
- Demo
- Internal tests
  - Brand Safety Report P...
  - Demonstration Report**

## Demonstration Report

Campaign length

13 DAYS

FILTERS

Dates

Include all

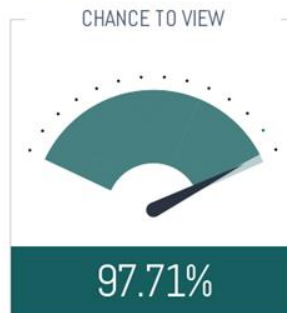
Domains

bakonyifoci.hu

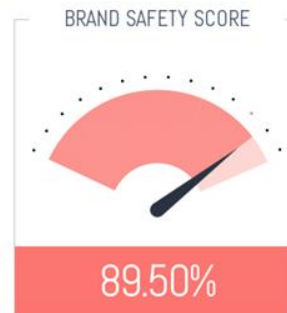
Ad fraud score



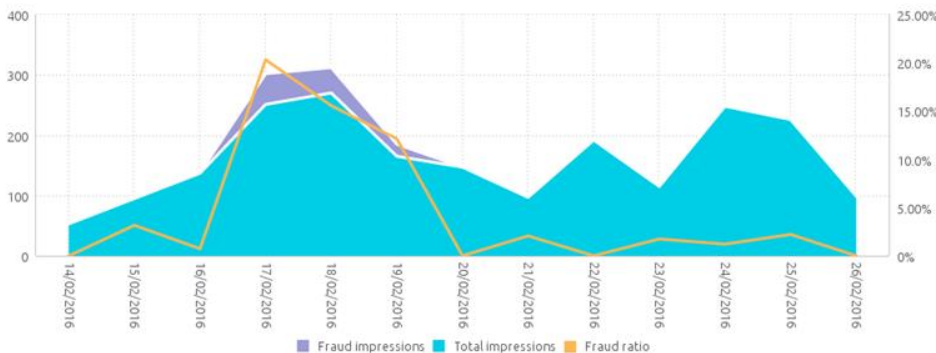
Chance To View Score



Brand Safety Score



Daily impressions



Total impressions

2,095 TOTAL IMPRESSIONS

Daily unique domains



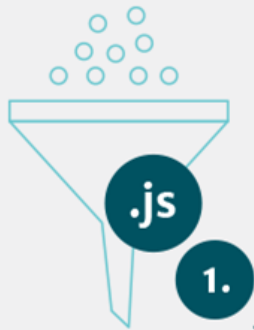
Unique domains

# DATA PLATFORM

---

# at Enbrite.ly

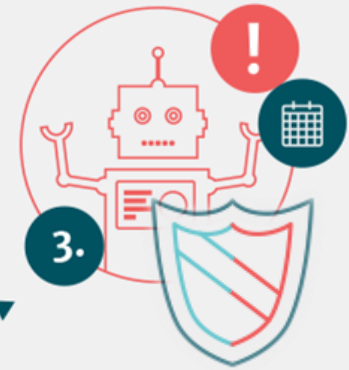
# What we do?



**DATA  
COLLECTION**

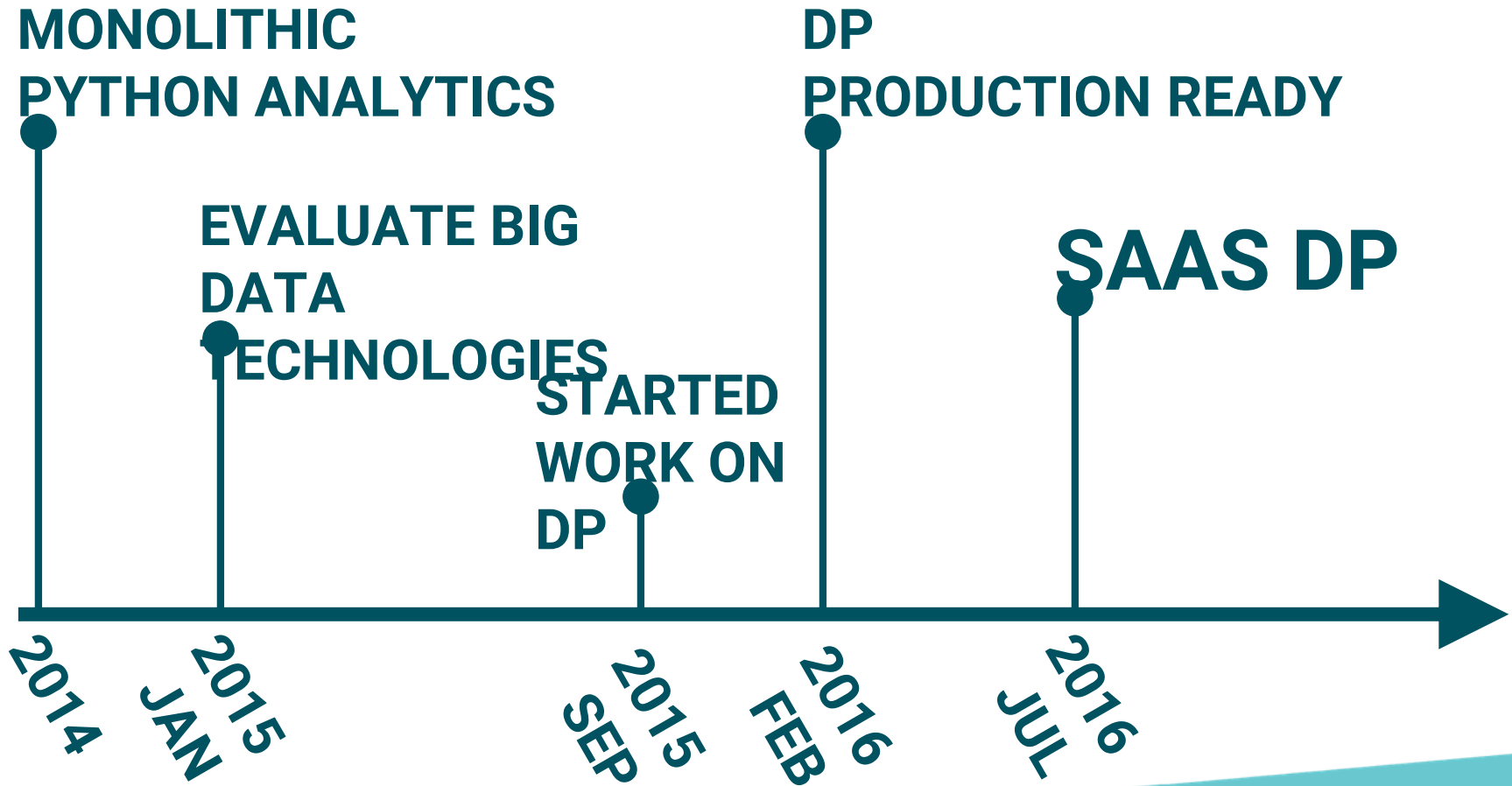


**ANALYZE  
DATA PROCESSION**



**ANTI FRAUD  
VIEWABILITY  
BRAND SAFETY  
REPORT + API**

# HOW WE GOT HERE?







# DATA COLLECTION

# The way to access log

## Click event attributes

(created by JS tracker)

```
{  
  "session_id": "spark_meetup_jsmmmq",  
  "timestamp": 1456080915621,  
  "type": "click"  
}
```

1.

2.

```
eyJzZXNzaW9uX2Ikljoic3BhcmtfbWVldHVwX2pzbW1tb3EiLCJ0aW1lc3RhbXAiOjE0NTYwODA5MTU2MjE5LnR5cGUiOiAiY2xpY2sifQo=
```

Access

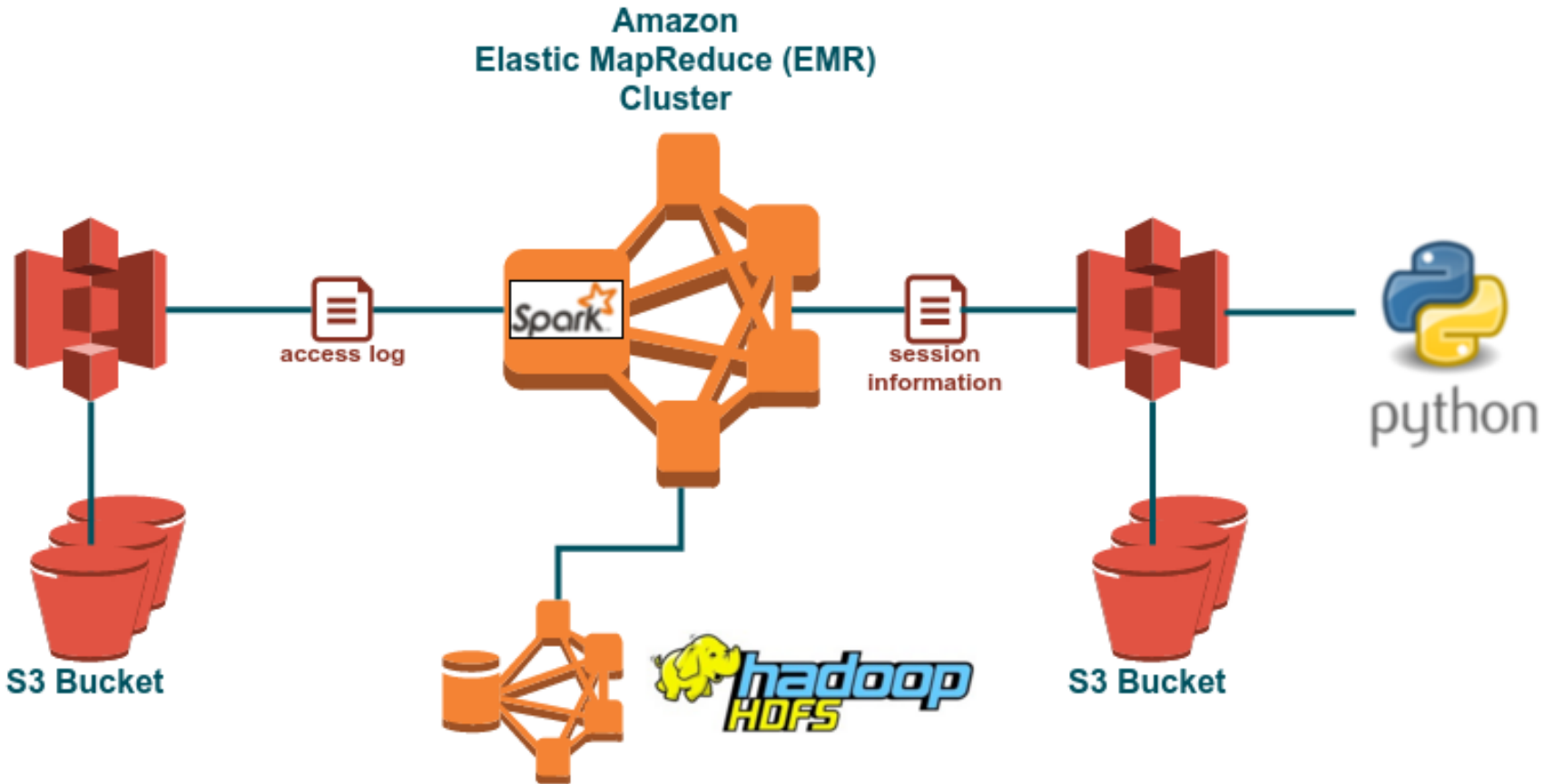
3.

log

format

```
TS CLIENT_IP STATUS "GET https://api.endpoint?event=eyJzZXNzaW9uX2Iklj..."
```

# DATA PROCESSING



# Spark

## TOOLS

- **0.5-2TB data processed daily**  
**1-10B rows**
- **Ad-hoc batch queries 20TB data**
- **20+ node cluster**
- **Spent 4 month optimizing it**

# Luigi

TOOLS

WORKFLOW ENGINE



Luigi + enbrite.ly extensions = **Gabo Luigi**

# LESSONS LEARNED

---

## LESSONS LEARNED

**YOU WILL SPEND A LOT  
OF TIME ON TOOLING**

# Value Checker

This page helps You check values on S3 to see if every variable is properly filled

WSID

Fields

Descriptives

Date

List Log Fi

- 2016-01-21
- 2016-01-21
- 2016-01-21
- 2016-01-21
- 2016-01-21
- 2016-01-21
- 2016-01-21
- 2016-01-21
- 2016-01-21
- 2016-01-21
- 2016-01-21\_00:00\_54.173.33.181
- 2016-01-21\_00:00\_54.173.33.181
- 2016-01-21\_00:00\_54.173.33.181
- 2016-01-21\_00:00\_54.173.33.181
- 2016-01-21\_00:00\_54.173.33.181
- 2016-01-21\_00:00\_54.173.33.181



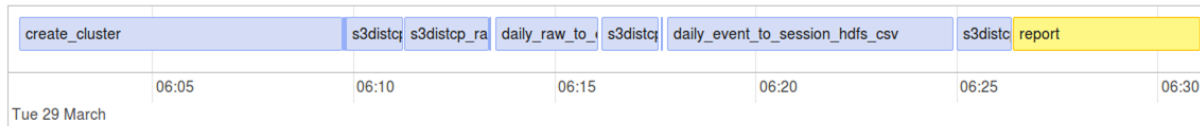
Check Selected Files



# Tools we created

GABO LUIGI

## Timeline



## Details

Show 15 entries

Search:

ID	Name	Type	StartTime	EndTime	Duration (h:m:s)	Status
12957	delete_hdfs_tear-down	DeleteHDFS	2016-03-29 06:31:07	2016-03-29 06:31:13	00:00:06	DONE
12956	delete_master_tear-down	DeleteMaster	2016-03-29 06:31:05	2016-03-29 06:31:07	00:00:02	DONE
12955	report	LogragateTask	2016-03-29 06:26:23	2016-03-29 06:31:04	00:04:41	DONE
12954	s3distcp_session_hdfs_to_s3	S3DistCp	2016-03-29 06:24:59	2016-03-29 06:26:22	00:01:23	DONE
12953	daily_event_to_session_hdfs_csv	SparkJavaTask	2016-03-29 06:17:48	2016-03-29 06:24:55	00:07:07	DONE
12952	events_on_hdfs	RawLog	2016-03-29 06:17:38	2016-03-29 06:17:41	00:00:03	DONE
12951	s3distcp_state_s3_to_hdfs	S3DistCp	2016-03-29 06:16:09	2016-03-29 06:17:35	00:01:26	DONE
12950	daily_raw_to_event_hdfs	SparkJavaTask	2016-03-29 06:13:31	2016-03-29 06:16:05	00:02:34	DONE
12949	copy_raw_to_event_config	FileCopy	2016-03-29 06:13:22	2016-03-29 06:13:24	00:00:02	DONE
12948	s3distcp_raw_s3_to_hdfs	S3DistCp	2016-03-29 06:11:15	2016-03-29 06:13:22	00:02:07	DONE
12947	s3distcp_raw_tomorrow_s3_to_hdfs	S3DistCp	2016-03-29 06:09:49	2016-03-29 06:11:14	00:01:25	DONE
12946	copy_jar	FileCopy	2016-03-29 06:09:47	2016-03-29 06:09:48	00:00:01	DONE
12945	copy_event_to_session_config	FileCopy	2016-03-29 06:09:43	2016-03-29 06:09:46	00:00:03	DONE
12944	create_cluster	CreateCluster	2016-03-29 06:01:41	2016-03-29 06:09:43	00:08:02	DONE

Showing 1 to 14 of 14 entries

Previous 1 Next

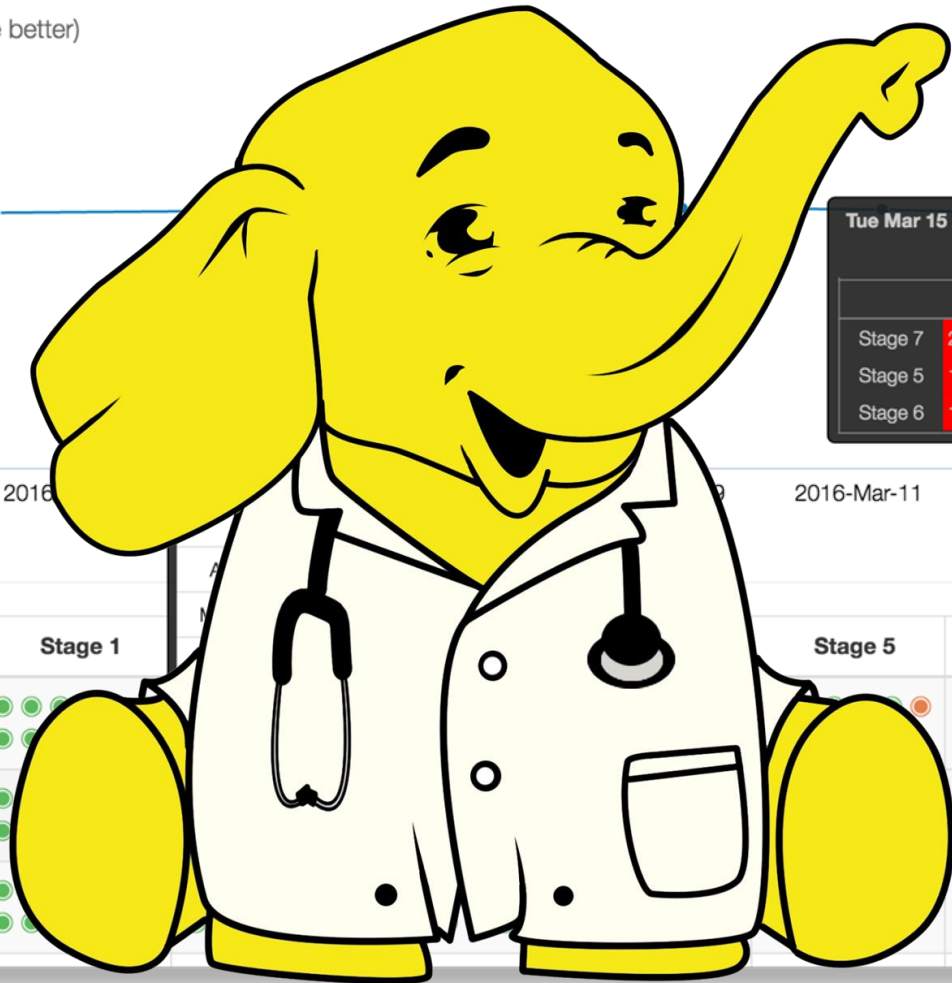
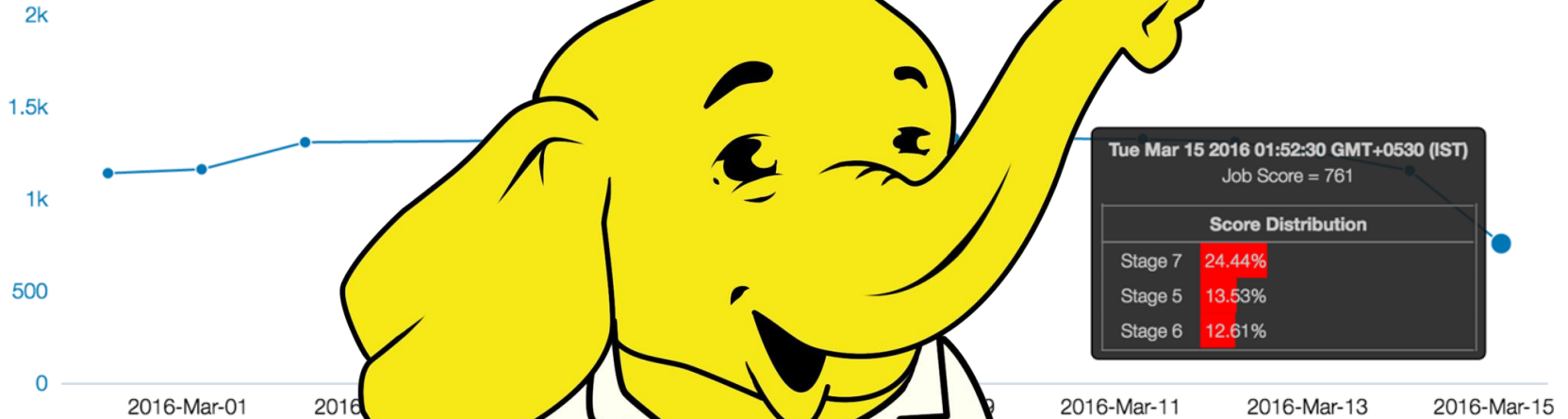
OPTIMIZATION

takes a

LOT OF TIME

Job History Results: <https://dev-azkaban:9000/manager?project=project-name&flow=flow-name&job=job-name>

Performance Score (Lower the better)



Job Executions	Stage 1	Stage 5	Stage 6	Stage 7
Mar 15, 2016 01:52 AM	●●●●●	●	●●●●●●●●●●	●●●●●●●●●●
Mar 14, 2016 02:20 AM	●●●●●		●●●●●●●●●●	●●●●●●●●●●
Mar 13, 2016 03:24 AM	●●●●●		●●●●●●●●●●	●●●●●●●●●●

# OPTIMIZATION NEVER ENDS

**AUTOMATE  
PERFORMANCE  
OPTIMIZATION**

# PERFORMANCE MEASUREMENTS

- **CLUSTER CONFIGURATION**
- **SPARK JOB CONFIGURATION**
- **DATA SET VARIATIONS**
- **IMPACT OF ALGORITHMS**

# PERFORMANCE MEASUREMENTS



















# MARATHON
















**DATA STORAGE IS THE  
BIGGEST  
OPTIMIZATION**




































# DON'T START WITH SCALA AND SPARK

**KEEP ANALYTICS CODE  
IN ONE  
REPOSITORY**



















- ▶  eventtosession
  -   Compactify
  -   DailyEventToSession
  -   DailyGeneralUrlParamsToEvent
  -   DailyRawToEvent
  -   EventToSession
  -   GeneralUrlParamsToEvent
  -   RawToEvent
  -  ◦ RawToEventConfig
  -   SessionSmartJoinRedirect
















- ▼  unit
  -   Aggregator
  -   CreateReports
  -   RecordAggregator
  -   RecordPreprocessing
  -   SessionAggragationWithGroupBy
  -   SessionAggregator
  -   SessionPostprocessing



















- ▶  eventtosession
  -   Compactify
  -   DailyEventToSession
  -   DailyGeneralUrlParamsToEvent
  -   DailyRawToEvent
  -   EventToSession
  -   GeneralUrlParamsToEvent
  -   RawToEvent
  -  ◦ RawToEventConfig
  -   SessionSmartJoinRedirect
















- ▼  unit
  -   Aggregator
  -   CreateReports
  -   RecordAggregator
  -   RecordPreprocessing
  -   SessionAggragationWithGroupBy
  -   SessionAggregator
  -   SessionPostprocessing

# STRUCTURE YOUR CODE

- ▶  eventtosession
  -   Compactify
  -   DailyEventToSession
  -   DailyGeneralUrlParamsToEvent
  -   DailyRawToEvent
  -   EventToSession
  -   GeneralUrlParamsToEvent
  -   RawToEvent
  -  ◦ RawToEventConfig
  -   SessionSmartJoinRedirect

- ▼  unit
  -   Aggregator
  -   CreateReports
  -   RecordAggregator
  -   RecordPreprocessing
  -   SessionAggragationWithGroupBy
  -   SessionAggregator
  -   SessionPostprocessing

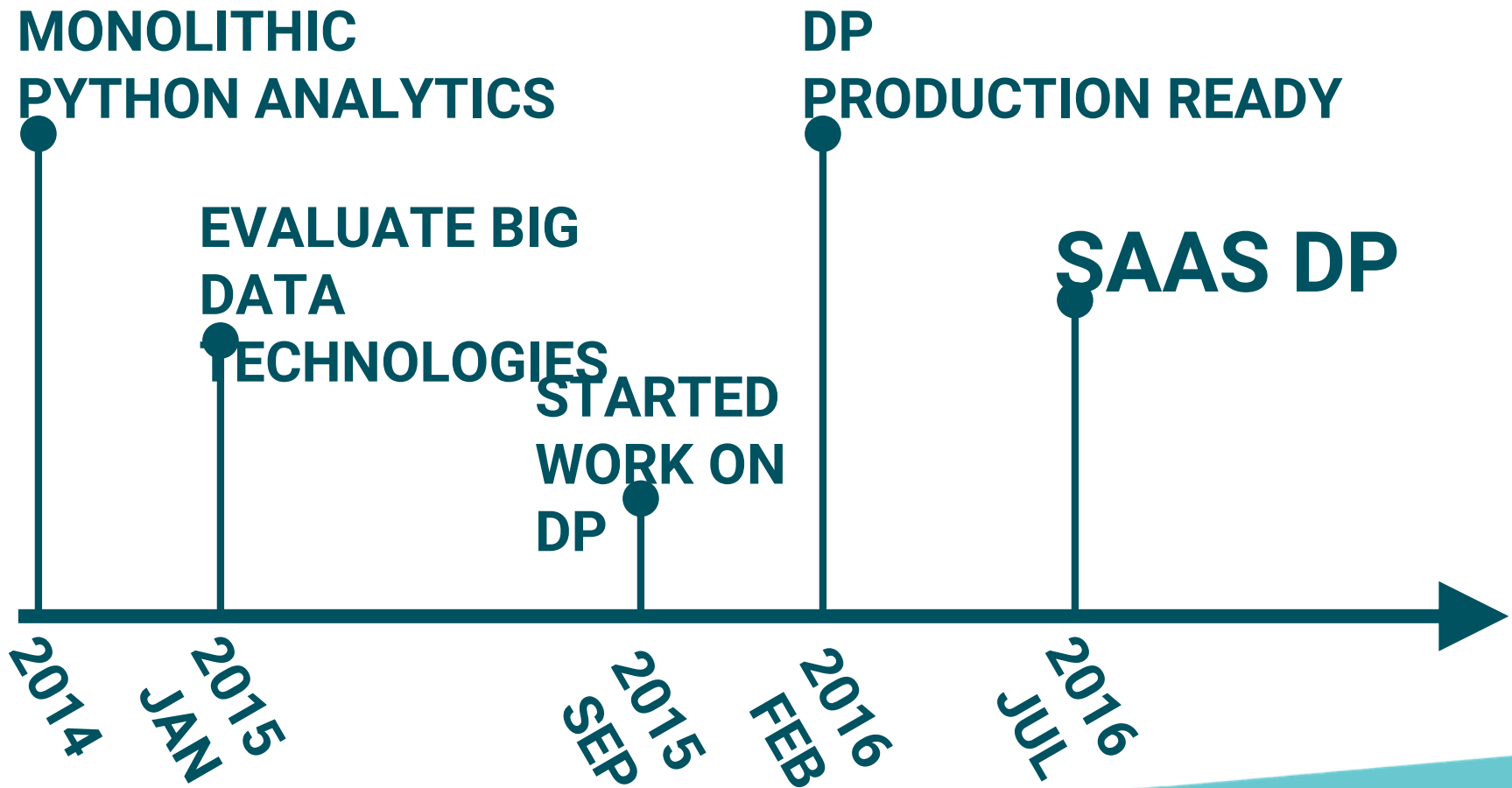
- ▶  eventtosession
  -   Compactify
  -   DailyEventToSession
  -   DailyGeneralUrlParamsToEvent
  -   DailyRawToEvent
  -   EventToSession
  -   GeneralUrlParamsToEvent
  -   RawToEvent
  -  ◦ RawToEventConfig
  -   SessionSmartJoinRedirect

- ▼  unit
  -   Aggregator
  -   CreateReports
  -   RecordAggregator
  -   RecordPreprocessing
  -   SessionAggragationWithGroupBy
  -   SessionAggregator
  -   SessionPostprocessing

# START WITH THE SMALLEST BIG DATA PROJECT



# HOW WE GOT HERE?



# CODE REUSE

# KNOWLEDGE REUSE

# Unified Data Processing Engine



**NOT EVERY USE CASE**  

---

**IS A SPARK USE-CASE**

# THANK YOU!

**MATE GULYAS**

*gulyasm@enbrite.ly*

*@gulyasm*  
*@enbritely*

