Modern Data Pipelines in AdTech - life in the trenches
Roksolana Diachuk

• Big Data Developer at Captify
• Diversity & Inclusion ambassador at Captify
• Women Who Code Kyiv Data Engineering Lead
• Speaker
Agenda

1. What is AdTech?
2. Data pipelines in AdTech
3. Practical examples
4. Historical data reprocessing
5. Conclusions
AdTech

AdTech methodologies deliver the right content at the right time to the right consumer.
What Captify does?

Captify’s technologies unite to collect, connect and categorise billions of real-time search events from 2.3 billion consumers.
Introducing Captify SENSE
The Search Intent Platform

Captify’s new cookieless enabled end-to-end platform puts Search Intelligence at your fingertips—powering advanced audience planning and instant activation to drive superior media performance and efficiency.

“Google trends on steroids”
Chris Ashworth, Head of Strategy

“If the ongoing pandemic has demonstrated anything, it’s that pre-packaged pre-COVID consumer data no longer applies”
Alison Shiff, Senior Editor
Data pipelines in AdTech

- Reporting
Data pipelines in AdTech

- Reporting
- Insights
Data pipelines in AdTech

• Reporting
• Insights
• Data costs attribution
Data pipelines in AdTech

- Reporting
- Insights
- Data costs attribution
- Users audiences building
Data pipelines in AdTech

• Reporting
• Insights
• Data costs attribution
• Users audiences building
• All kinds of data processing/storage
Reporting

Data provider

Transformer

Loader
Data ingestion

Amazon S3 → CSV → Parquet → Apache Spark
Data ingestion

Amazon S3 → S3 lister → CSV → Parquet → 
Protobuf → Schema → Dates → Parsing → Apache Spark
Data loading
Data loading

Apache Spark

Schema definition

First upload vs scheduled ones

Cloudera Impala

Metadata handling

HDFS
Challenges

• Diverse data types
Challenges

• Diverse data types
• Time dependency
Challenges

- Diverse data types
- Time dependency
- External data storage
Challenges

- Diverse data types
- Time dependency
- External data storage
- Constant connection with end users
Data costs attribution
Data costs attribution

Log-level data → Ingestor → Mapper → Transformer

Data costs calculator
Data costs attribution

Amazon S3 → Apache Spark → HDFS

PostgreSQL
Attribution data source

Amazon S3

- Standard feed
- Segment feed
Data costs attribution

Amazon S3 → Ingestion → Mapping → Transformation

PostgreSQL

HDFS

Data costs calculator
Data costs attribution

- Amazon S3
  - S3 lister
  - Ingestion
- Unified schema
- Mapping
- Transformation
  - Feeds join
  - Data costs
- PostgreSQL
- HDFS
- Data costs calculator
Challenges

• Processing and storing really large data volumes (!)
Challenges

• Processing and storing really large data volumes (!)
• Failures handling
Challenges

- Processing and storing really large data volumes (!)
- Failures handling
- Historical data reprocessing
Historical data reprocessing
Business use case
Attribution pipeline
Reprocessing mechanism

standardfeed.transformer.Config.feedPeriod: “P30D”
standardfeed.transformer.Config.startDateTime: 2022-03-01T00:00

val minTime = currentDay.minus(config.feedPeriod)

listFiles.filter(file => file.eventDateTime isAfter minTime)
Reprocessing

Production version

Reprocessed version

HDFS

is_impression=true

is_impression=false

is_impression=true

is_impression=false
Production version

Reprocessed version

HDFS

cloudera
IMPALA

attribution_prod

attribution_reprocessed
Reporting

HDFS -> Spark -> Cloudera

Dropped partitions -> Reprocessed table
Future with Delta lake

Schema enforcement

Time travel

Keeping track of changes
Future with Delta lake

Parquet files => Delta files

Spark tables => Delta tables

... Leveraging data versions through Delta tables history

Vacuum unsuitable data
Challenges

• Computing resources
Challenges

• Computing resources
• Speed of processing
Challenges

- Computing resources
- Speed of processing
- Complexity
Challenges

• Computing resources
• Speed of processing
• Complexity
• High cost of the errors
Conclusions
Conclusions

1. AdTech is an exciting domain for big data
Conclusions

1. AdTech is an exciting domain for big data

2. There is more than one approach to leveraging data
Conclusions

1. AdTech is an exciting domain for big data

2. There is more than one approach to leveraging data

3. There is always a room for improvement
My contact info

dead_flowers22
roksolana-d
roksolanadiachuk
roksolanad
Stand With Ukraine