Streaming datasets for Personalization

Shriya Arora

Senior Data Engineer
Personalization Analytics

NETFLIX

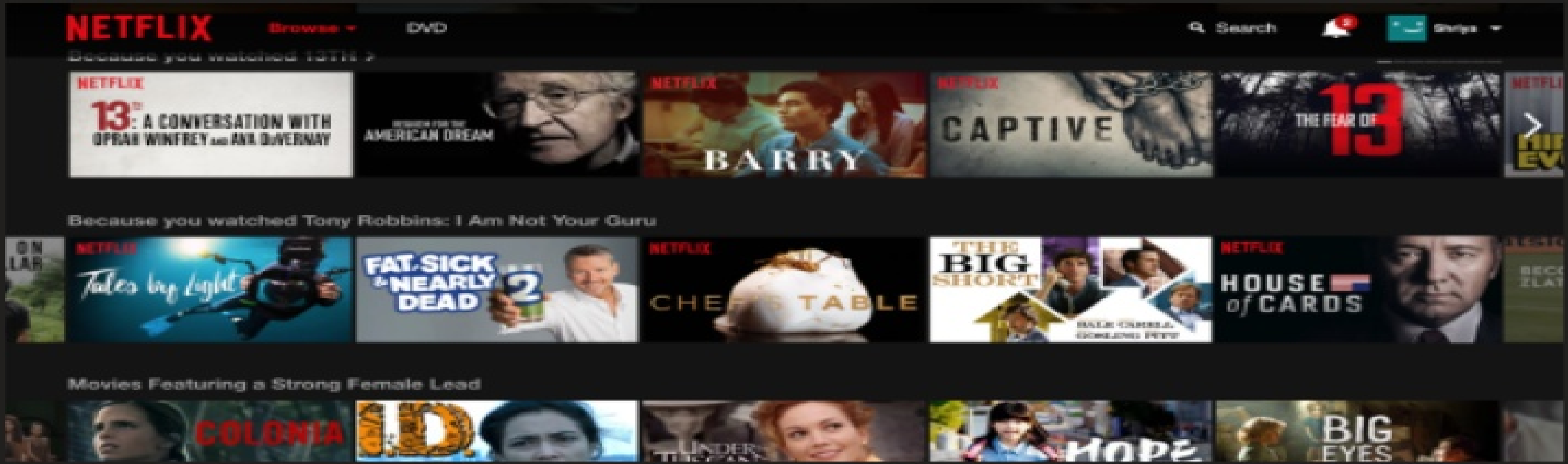
What is Netflix's Mission?

Entertaining you by allowing you to stream content anywhere, anytime

NETFLIX

What is Netflix's Mission?

Entertaining you by allowing you to stream personalized content anywhere, anytime

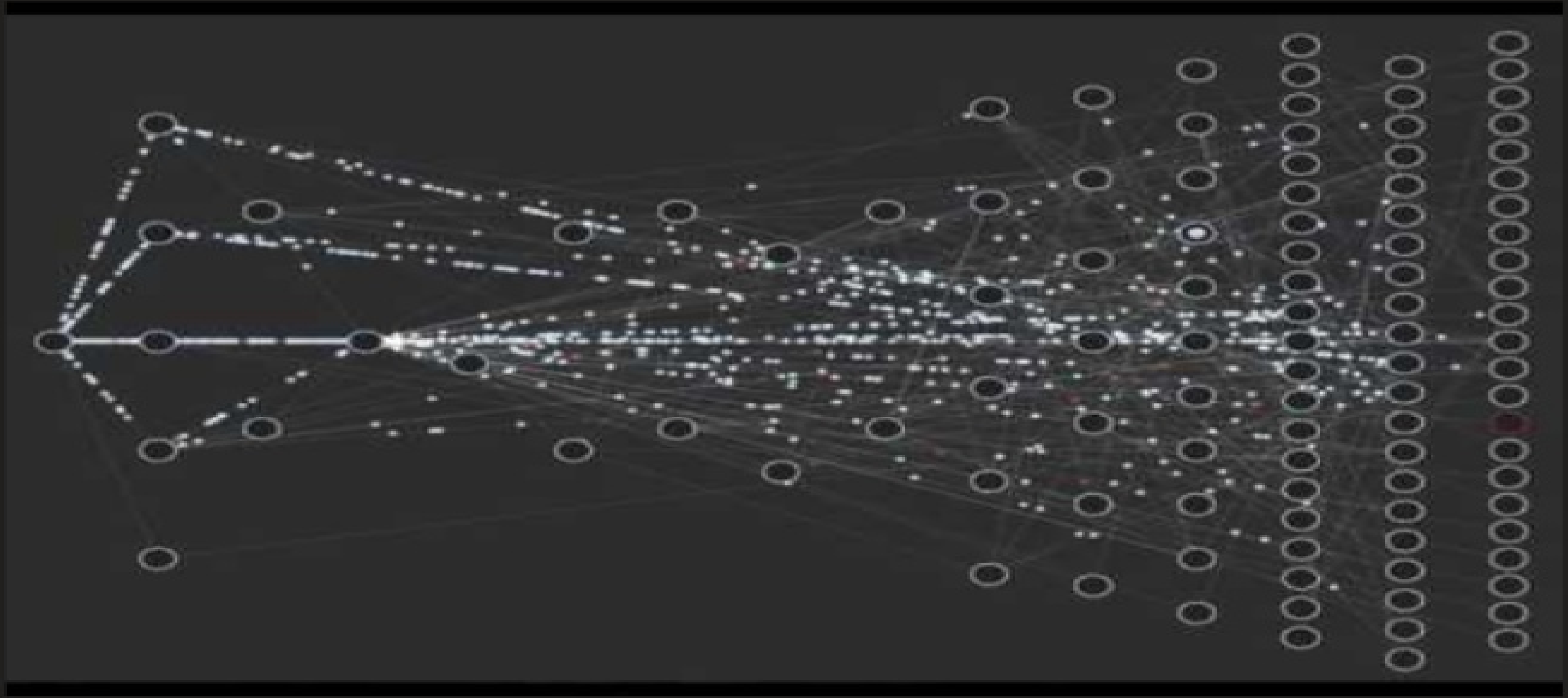


NETFLIX

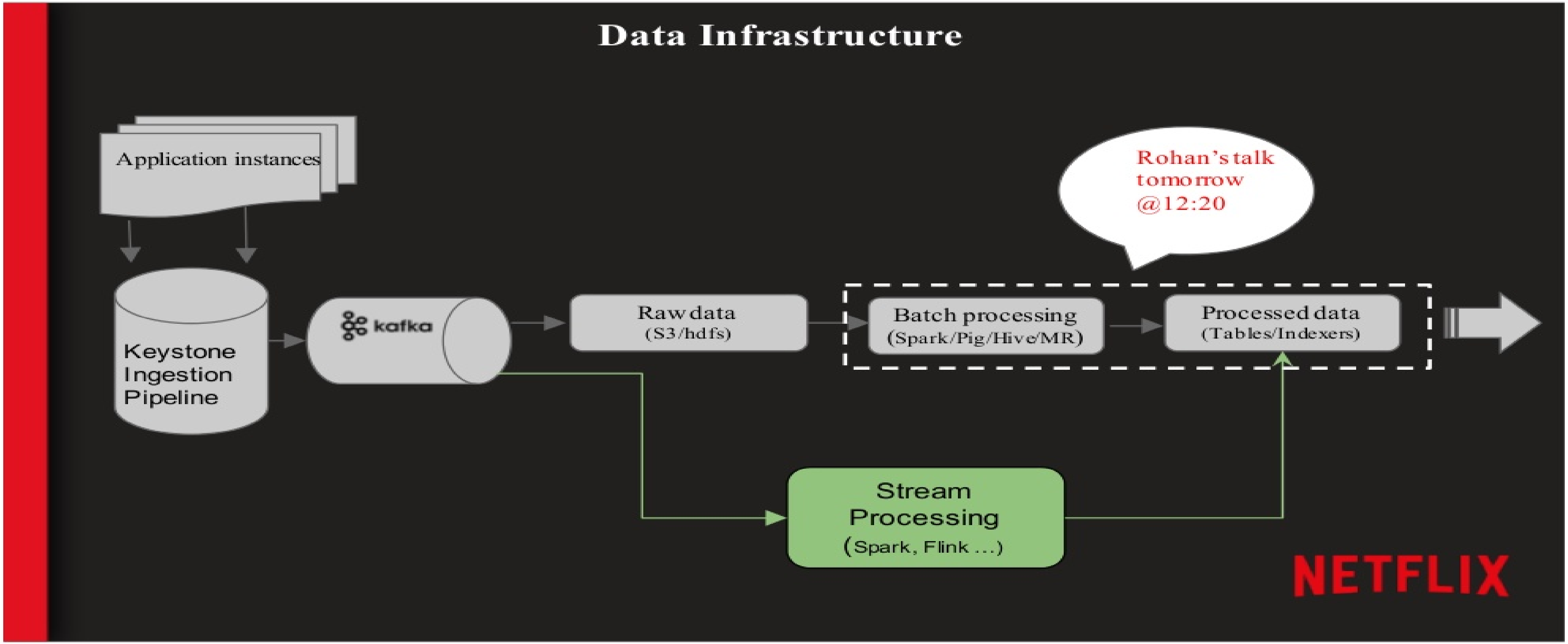
How much data do we process to have a personalized Netflix for everyone?

- 93M+ active members
- 125M hours/ day
- 190 countries with unique catalogs
- 450B unique events/day
- 600+ Kafka topics

A SERIES OF PLAYBACK EVENTS



NETFLIX


Data Infrastructure
Application instances
Keystone Ingestion Pipeline
kafka
Raw data (S3/hdfs)
Batch processing (Spark/Pig/Hive/MR)
Processed data (Tables/Indexers)
Rohan's talk tomorrow @12:20
Stream Processing (Spark, Flink ...)
NETFLIX

Our problem: Using user plays for feature generation, discovery, clustering ..

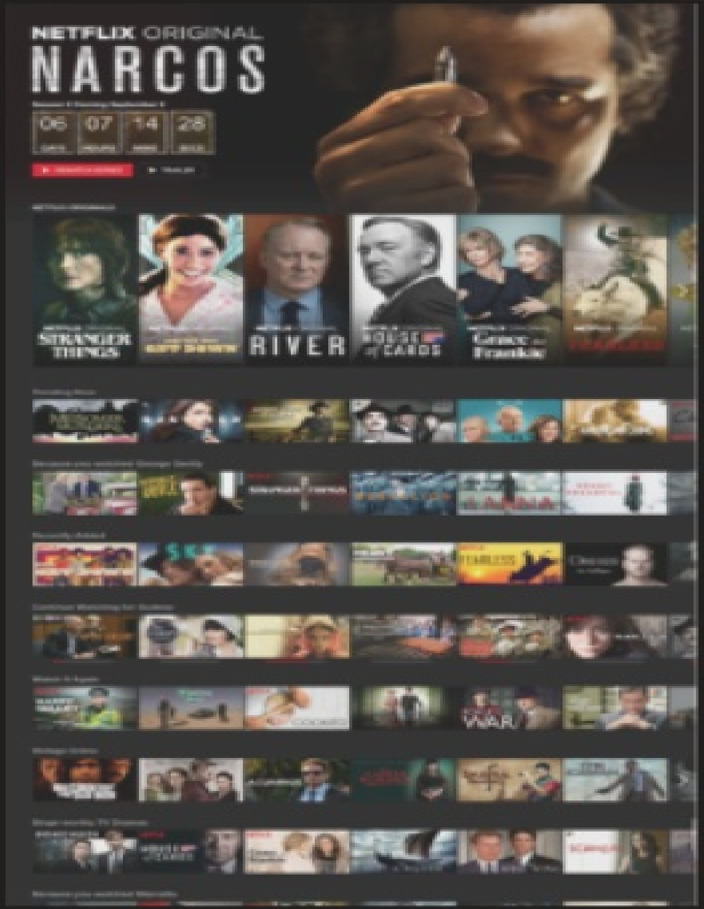
NETFLIX

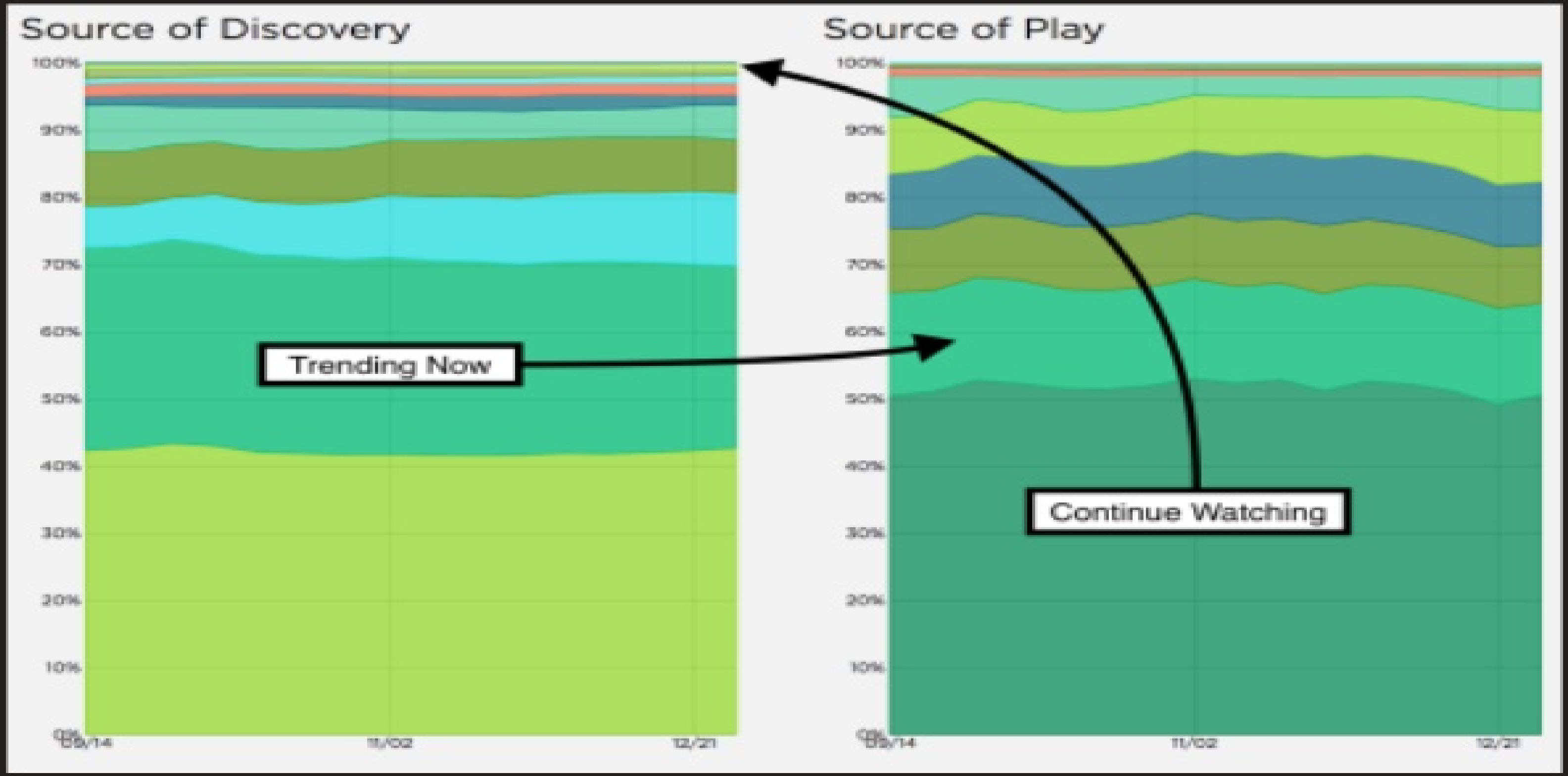
Why have data later when you can have it now?

- **Business Wins**
  - Algorithms can be trained with the latest + greatest data
  - Enhances research
  - Creates opportunity for new types of algorithms
- **Technical Wins**
  - Save on storage costs
  - Avoid long running jobs

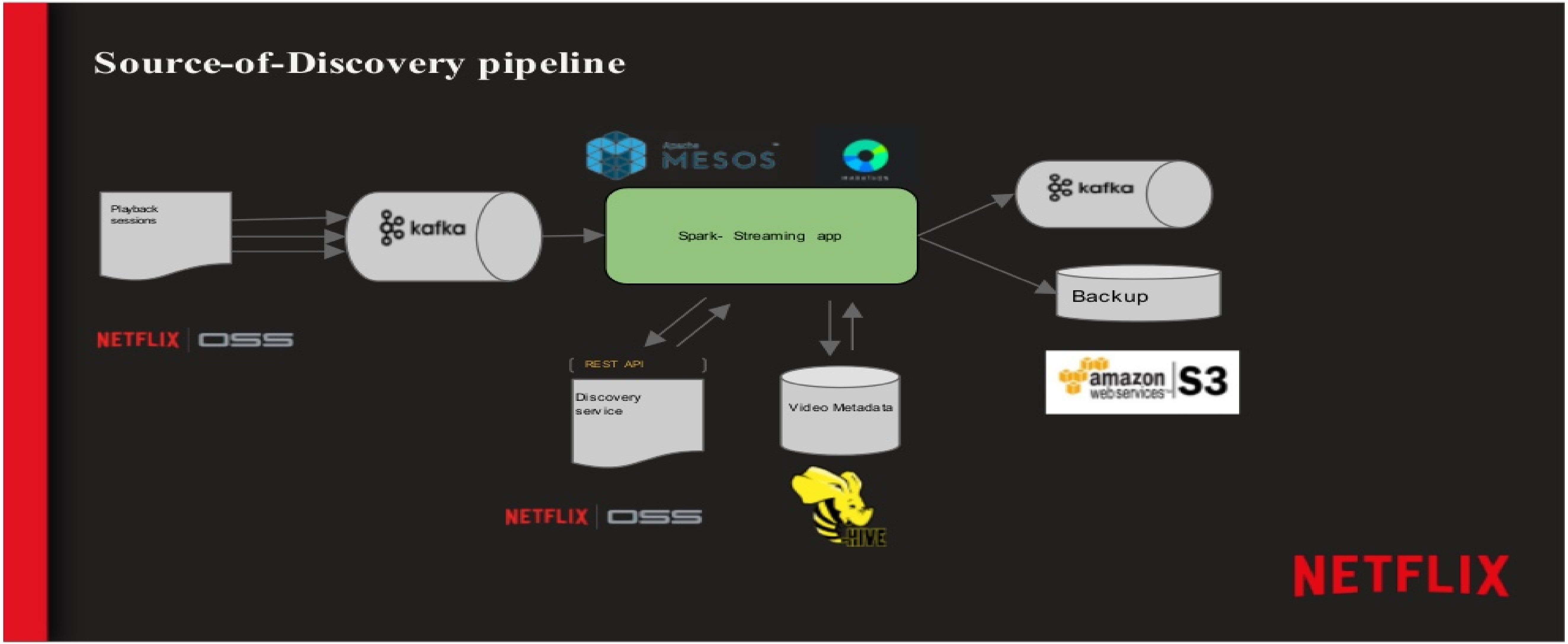
NETFLIX

Source of Discovery / Source of Play

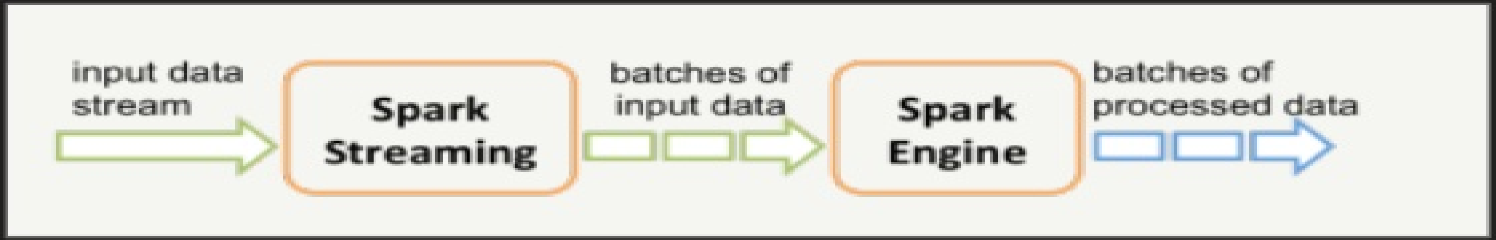




NETFLIX


Source-of-Discovery pipeline
Apache MESOS
Playback sessions
kafka
Spark- Streaming app
kafka
Backup
NETFLIX | OSS
REST API
Discovery service
Video Metadata
NETFLIX | OSS
amazon web services S3
HIVE
NETFLIX

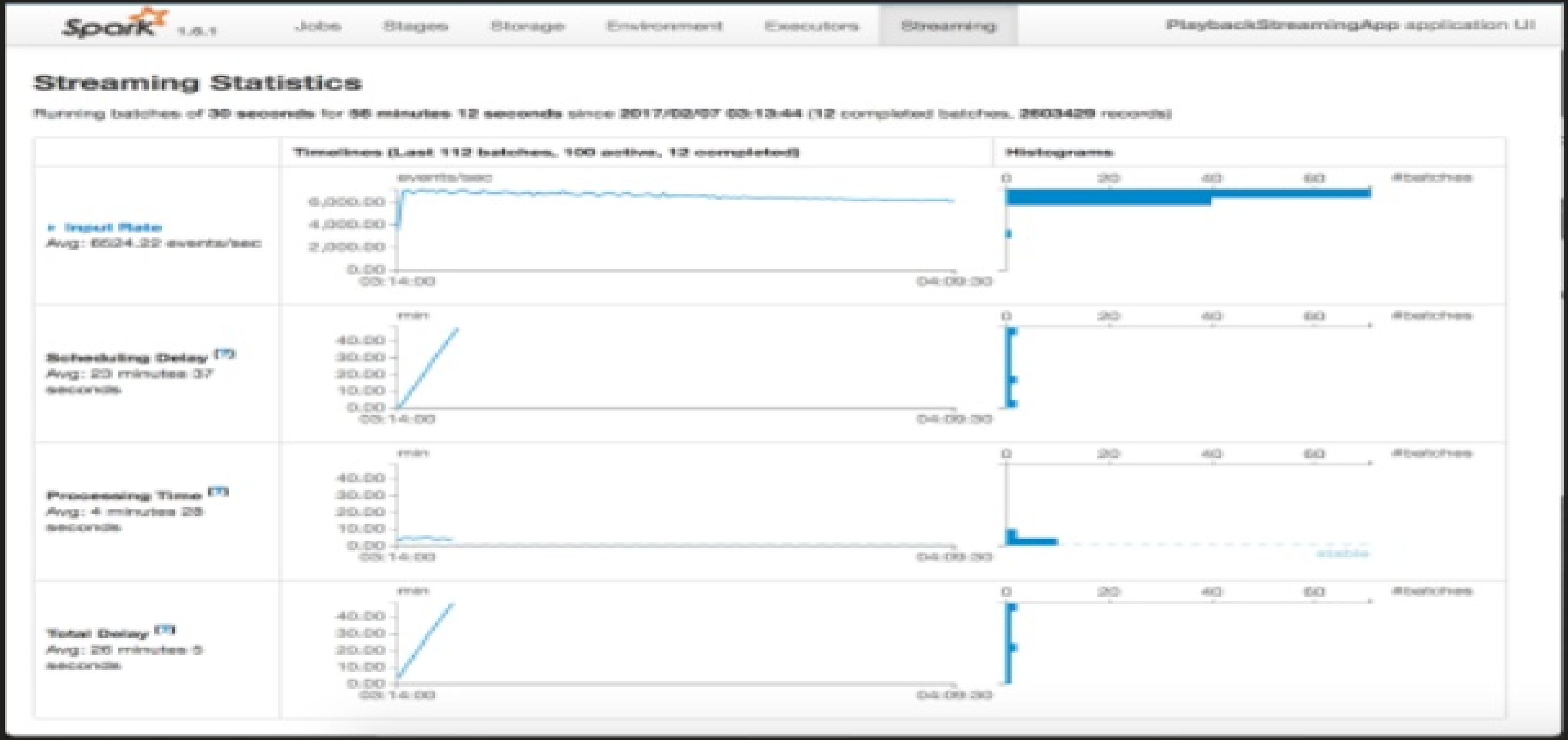
Spark Streaming



- Needs a StreamingContext and a batch duration
- Data received in DStreams, which are easily converted to RDDs
- Support all fundamental RDD transformations and operations
- Time-based windowing
- Checkpointing support for resilience to failures
- Deployment

NETFLIX

Performance tuning your Spark streaming application



NETFLIX

Performance tuning your Spark streaming application

- Choice of micro-batch interval
  - The most important parameter
- Cluster memory
  - Large batch intervals need more memory
- Parallelism
  - DStreams naturally partitioned to Kafka partitions
  - Repartition can help with increased parallelism at the cost of shuffle
- # of CPUs
  - <= number of tasks
  - Depends on how computationally intensive your processing is

Performance tuning your Spark streaming application

NETFLIX

Challenges with Spark

- Not a 'pure' event streaming system
  - Minimum latency of batch interval
  - Un-intuitive to design for a stream-only world

- Choice of batch interval is a little too critical
  - Everything can go wrong, if you choose this wrong
  - Build-up of scheduling delay can lead to data loss

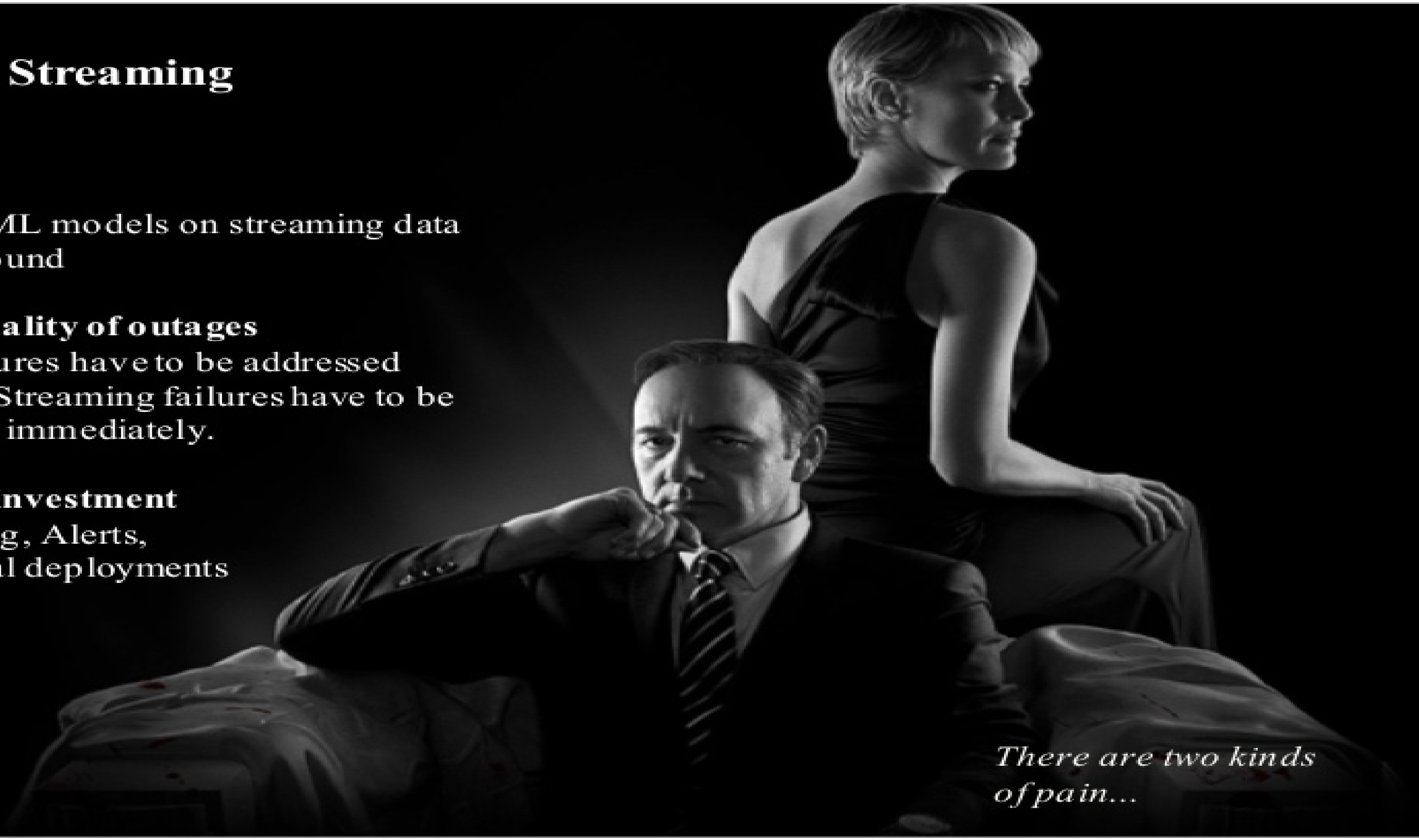
- Only time-based windowing*
  - Cannot be used to solve session-stitching use cases, or trigger based event aggregations

*** I used 1.6.1**

NETFLIX

Challenges with Streaming

- **Pioneer Tax**
  - Training ML models on streaming data is new ground
- **Increased criticality of outages**
  - Batch failures have to be addressed urgently, Streaming failures have to be addressed immediately.
- **Infrastructure investment**
  - Monitoring, Alerts,
  - Non-trivial deployments



Questions?

Stay in touch!

@NetflixData

NETFLIX