# HUNTING MALICIOUS CONTENT WITH SPARK

Perttu Ranta-aho
Ville Lindfors

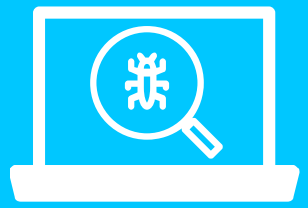
F-Secure

# F-SECURE

We offer our users the power to surf invisibly, securely store and share their personal data, and remain safe from online threats

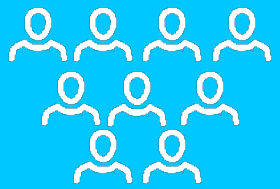
## SECURITY & PRIVACY

Founded in 1996, over 20 years in information security



## GLOBAL USERBASE

23 million users worldwide , through operator partners & direct consumers



## BASED IN FINLAND

European base, global operations

## SECURITY LABS

In-house teams ensure security for all our services across all platforms

# WHAT WE WANTED

- A platform for running data mining and analysis jobs
- Most common data sources easily available: Files (json, csv, etc.), PostgreSQL, Cassandra
- Tasks can write results to the data repository itself
- Possibility to use both high level languages (SQL/HiveQL) and real programming languages (Python/Java/Scala)

F-Secure

# WHAT WE DID

- ✓ Basic map-reduce jobs to get better touch what we have in our databases.
- ✓ Boost retraining of old ML-algorithms
- ✓ Started to replace some of our custom tool sets with MLlib & Spark
- ✓ Building up classifications of certain clusters of the web

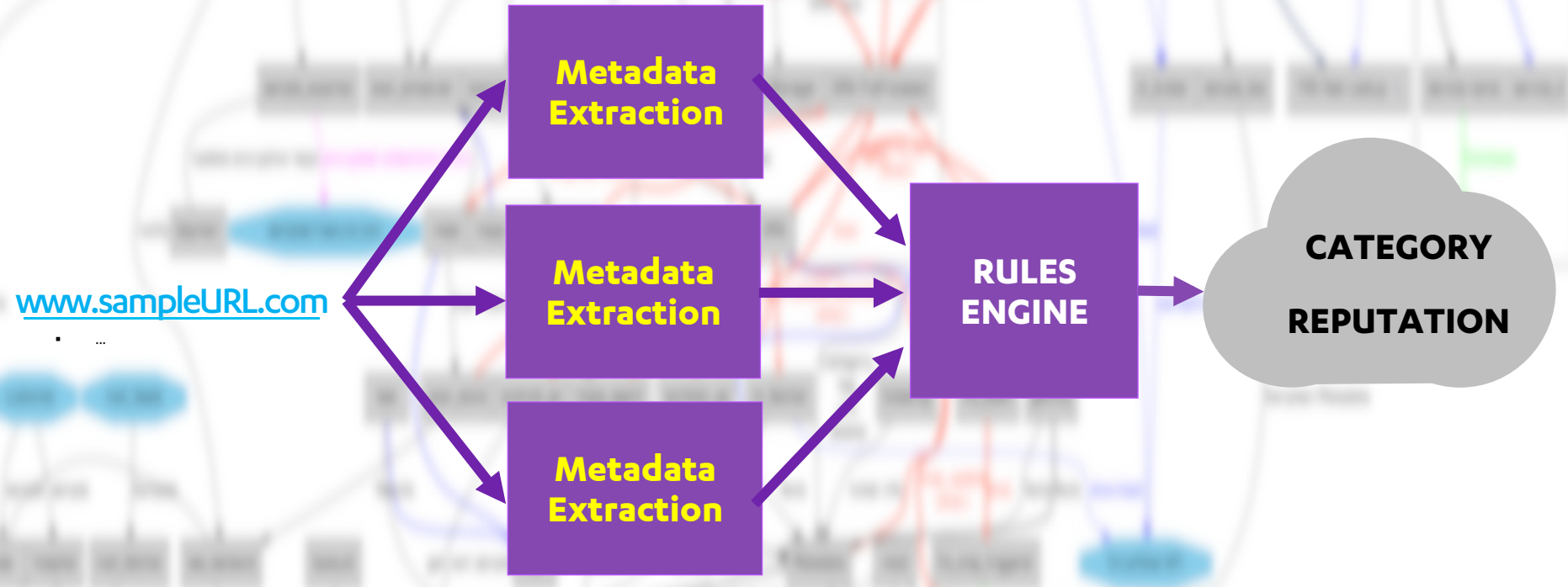
F-Secure

# CLIENT UPSTREAM

Statistics collected from clients to improve performance & coverage;
ANONYMIZED to preserve privacy

- Blocked infections
- Beta detections
- Hit counts
- Crashes
- …

# OUR FLOW


www.sampleURL.com
...
Metadata Extraction
Metadata Extraction
Metadata Extraction
RULES ENGINE
CATEGORY
REPUTATION

# WHAT'S GREAT

**PySpark!**

**Speed**

Fast to write,
fast to execute

**Ease of use**

Easy to start coding,
interactive shell!

**Batteries Included**

Streaming MLlib, Shark

**Upcoming**

Cassandra integration

F-Secure

# CHALLENGES

**Adaptation time**
Even good tools need adaptation time

**Tuning performance**
OutOfMemory, etc.

**Random issues**
PySpark crashes with bigger amounts of data, without clear reason

**Learning curve**
Fast to start but takes time to master

F-Secure

# NEXT STEPS

| EXPANSION | STREAMING | SHARK |
| --- | --- | --- |
| More users | Event processing | |
| More use cases | Anomaly Detection | |

# SWITCH ON FREEDOM

F-Secure