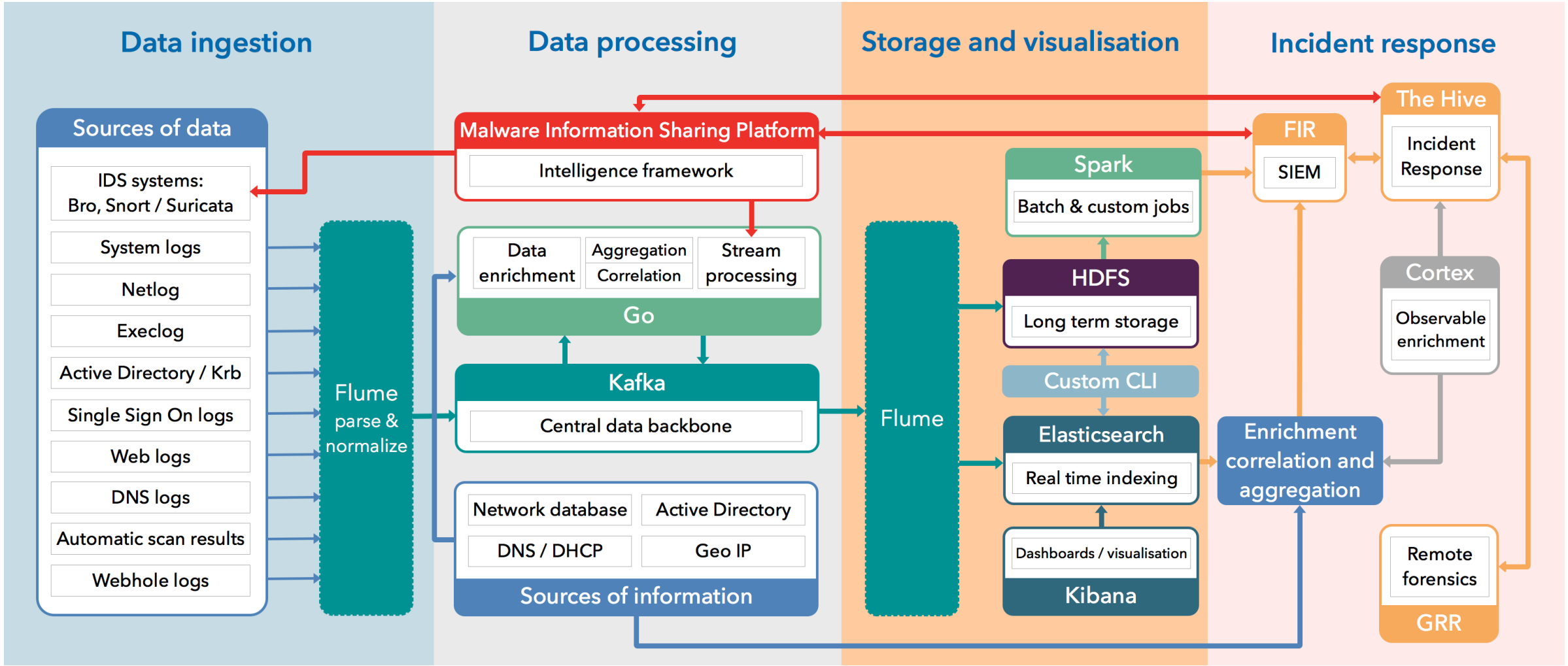
European Organization for Particle Physics

Exploring the frontiers of knowledge

CERN'S COMPUTER SECURITY OPERATIONS CENTRE

STATUS UPDATE

SYSTEM ARCHITECTURE



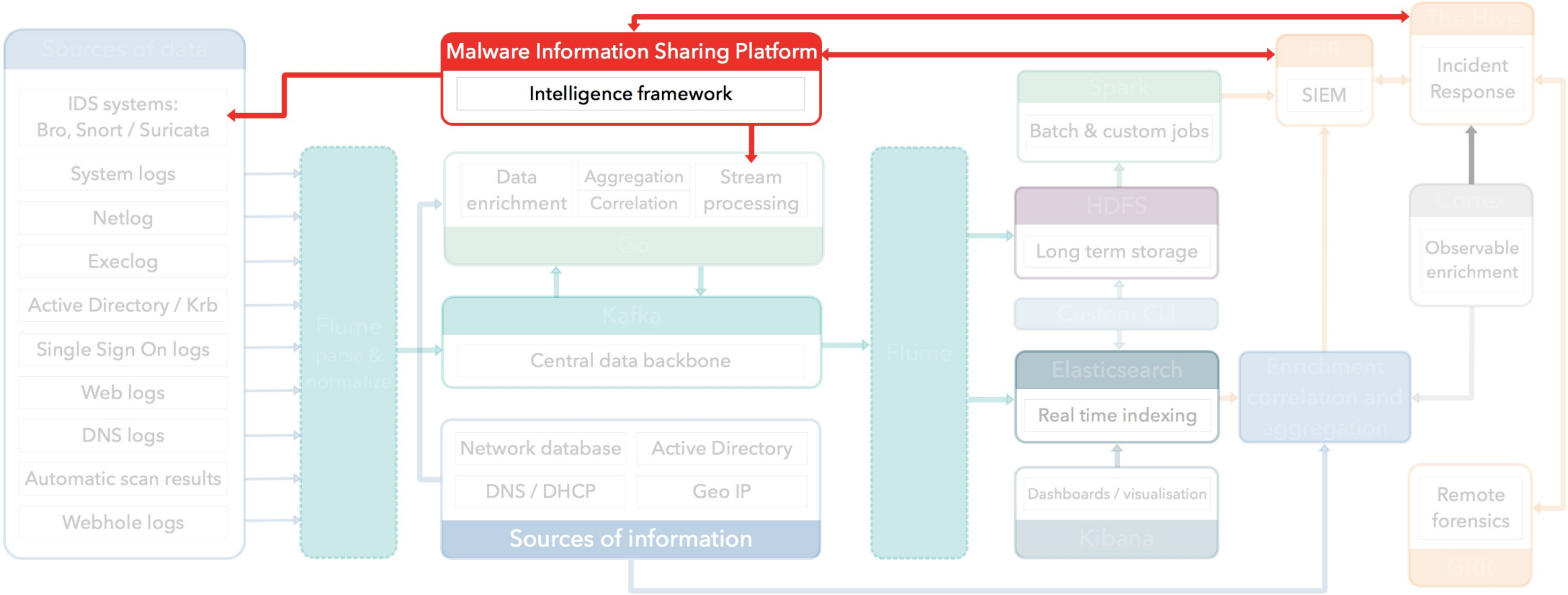
TECHNOLOGY STACK USED

- **Telemetry Capture Layer:** **Apache Flume**
- **Data Bus (Transport):** **Apache Kafka**
- **Analytics:** **Go**
- **Long-Term Data Store:** **Hadoop HDFS**
- **Real-Time Index & Search:** **Elasticsearch**
- **Visualisation:** **Kibana & CLI**
- **Intrusion Detection:** **Bro (Zeek) & Snort**
- **Web frontends:** **OpenShift**

DATA INGESTION RATES (1-7 FEB 2018)

- **Network (Bro / Zeek):**
  - **1078 GB / day in HDFS (raw json)**
  - **761 GB / day in ES**
  - **2.3 billion events / day**
- **System (other):**
  - **451 GB / day in HDFS (raw json)**
  - **256 GB / day in ES**
  - **1.1 billion events / day**

THREAT INTELLIGENCE

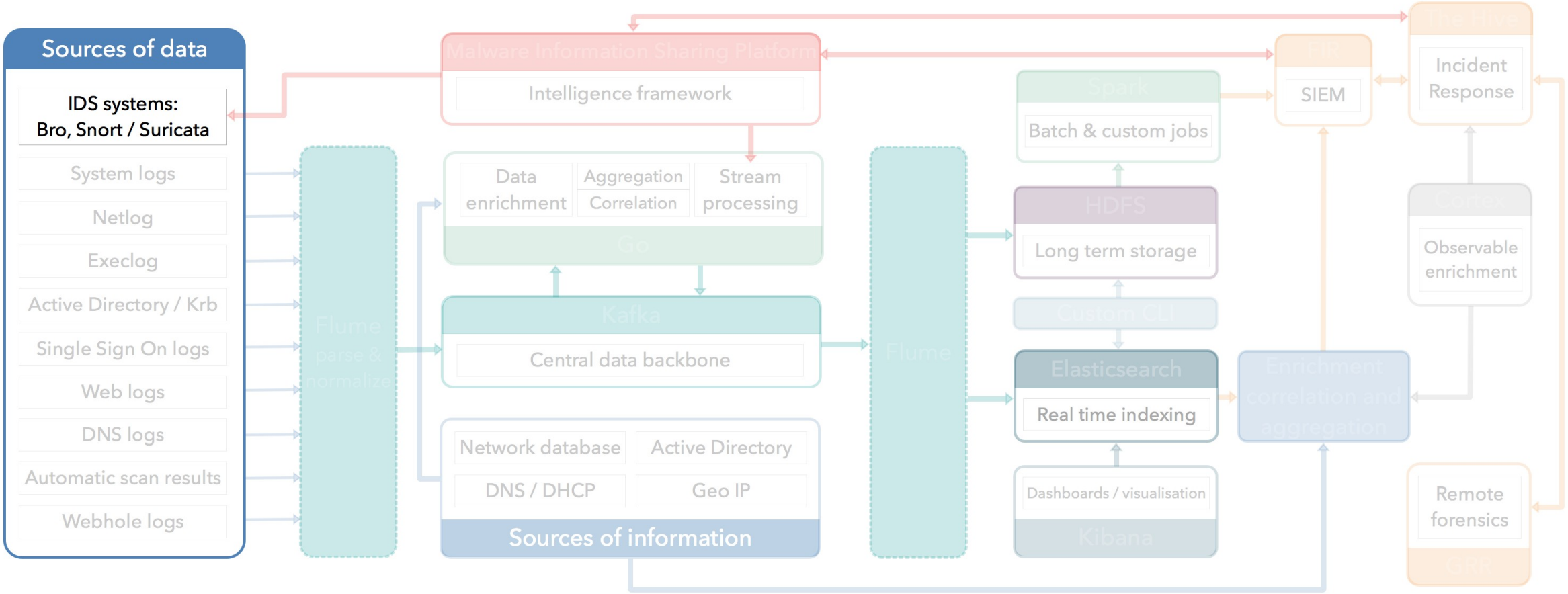


THREAT INTELLIGENCE

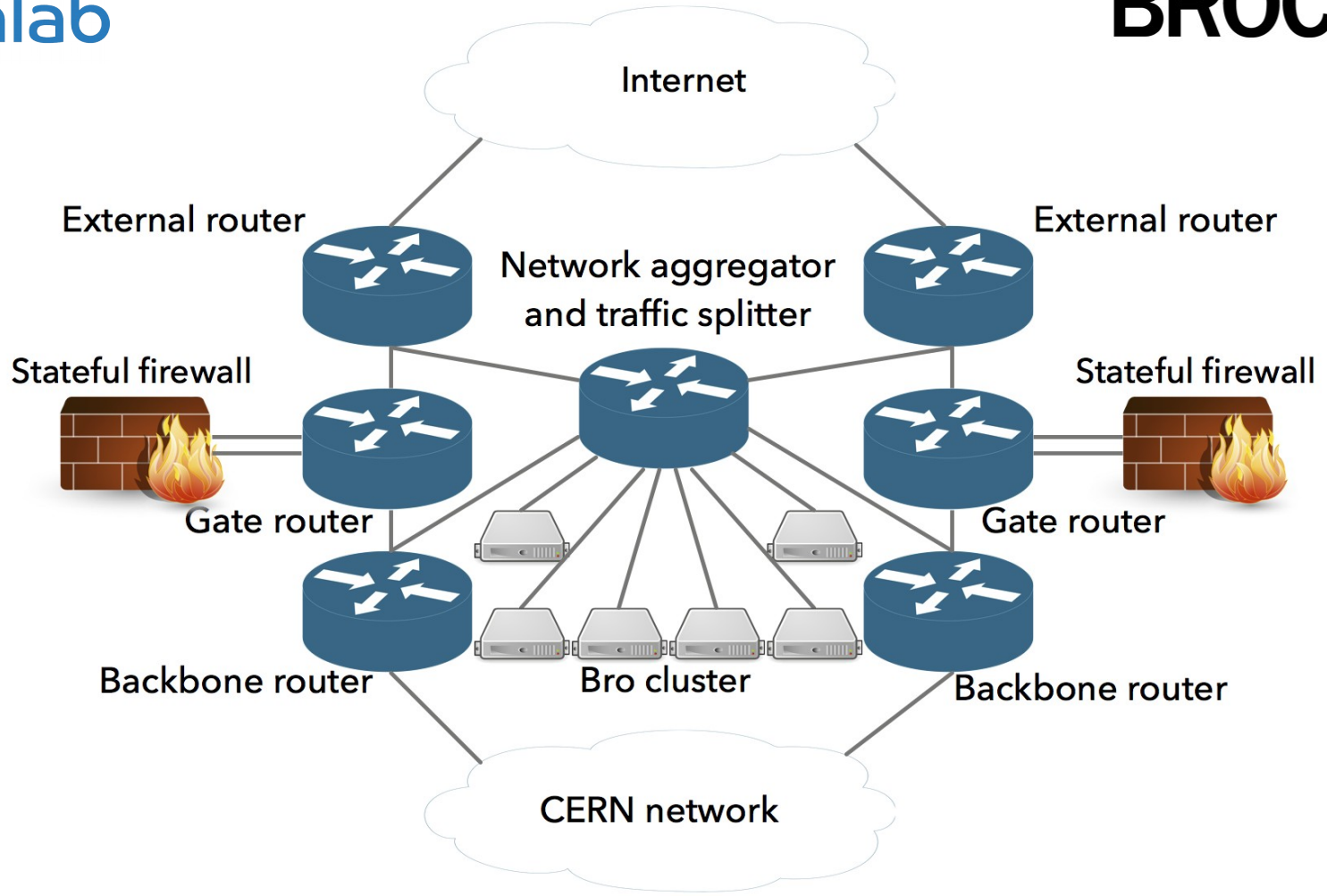
- Malware Information Sharing Platform (MISP) as the sole threat intelligence platform at CERN
  - Automatic sharing of intelligence data with trusted peers

- CERN is currently operating 4 different instances:
  - Main CERN instance (> 1.1 M IoCs)
  - Worldwide LHC Computing Grid (WLCG) central MISP instance (>600 K IoCs)
  - Development MISP instance used for MISP development (CERN is an active contributor) and for validating new MISP releases
  - Special purpose MISP instance

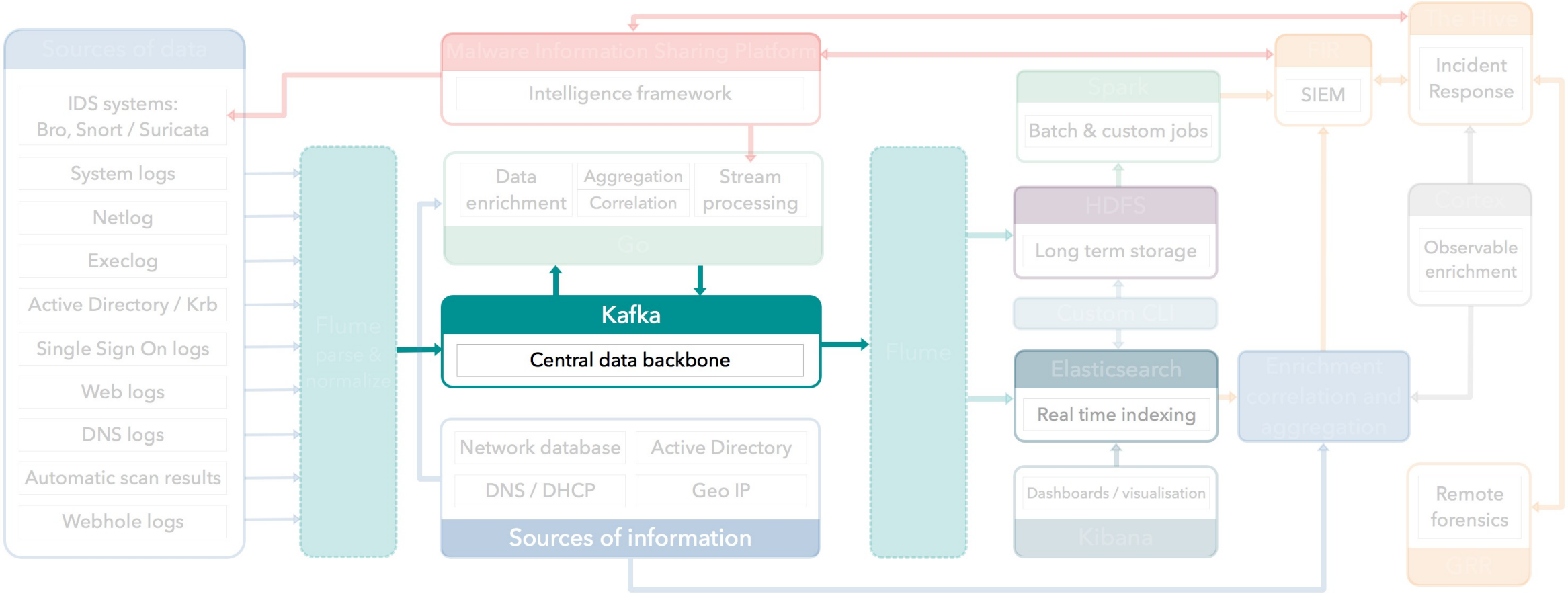
NETWORK BASED INTRUSION DETECTION



NETWORK TRAFFIC AGGREGATOR AND SPLITTER



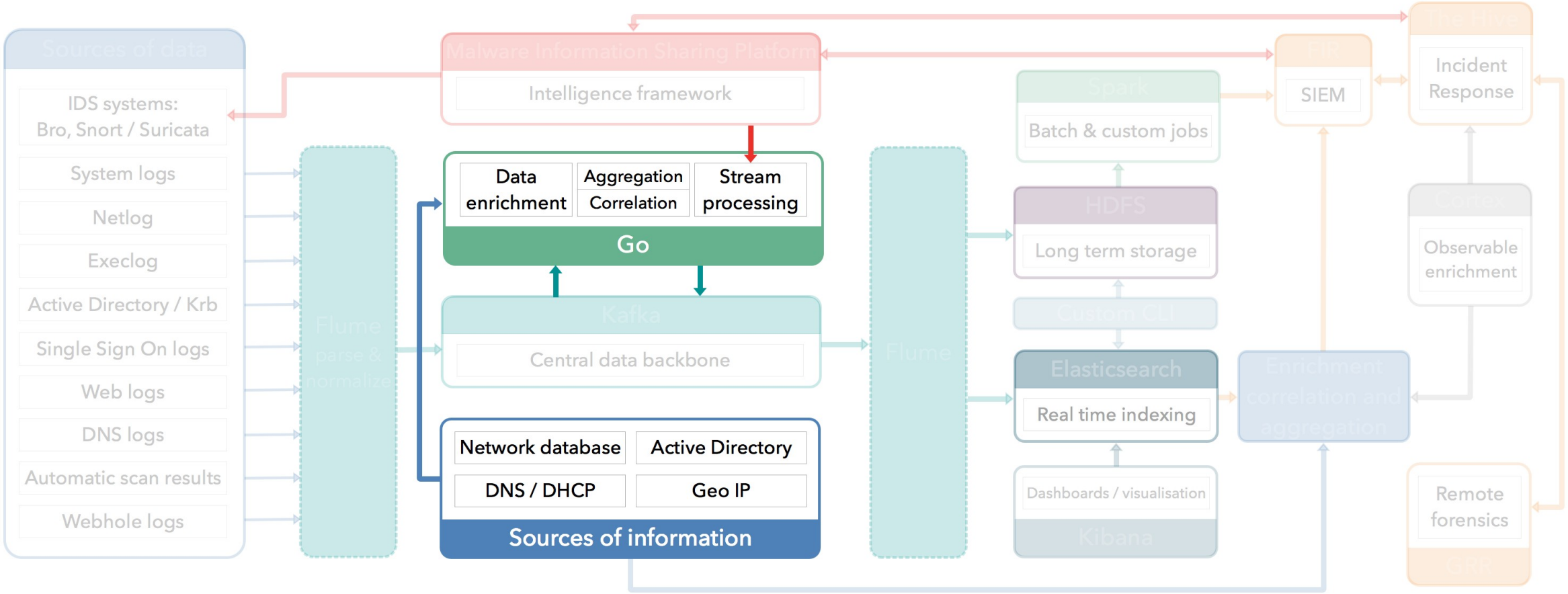
KAFKA DATA BACKBONE



KAFKA DATA BACKBONE

- New Kafka cluster
- 6 Kafka brokers, 3 Zookeeper nodes
  - 70,000 messages / sec on average
  - 72 hours retention period
  - Replication factor of 3
  - Data compressed using snappy

INLINE PROCESSING



INLINE PROCESSING

- Custom code written in golang
  - Jobs launched and monitored using Nomad
  - Running distributed on Nomad clients
- Data ingested from Kafka
- Types of jobs:
  - Data enrichment:
    - DNS (forward and reverse DNS resolutions)
    - GeoIP
  - Intrusion detection:
    - Based on IoCs from MISP
    - Custom, advanced rules
  - Monitoring
  - More to come

DATA ENRICHMENT

- Very fast, not guaranteed to be 100% accurate

- DNS resolution
  - Golang routines: highly asynchronous
  - ~1-3 sec delay for entries that can not be resolved
  - Filtering what messages to enrich

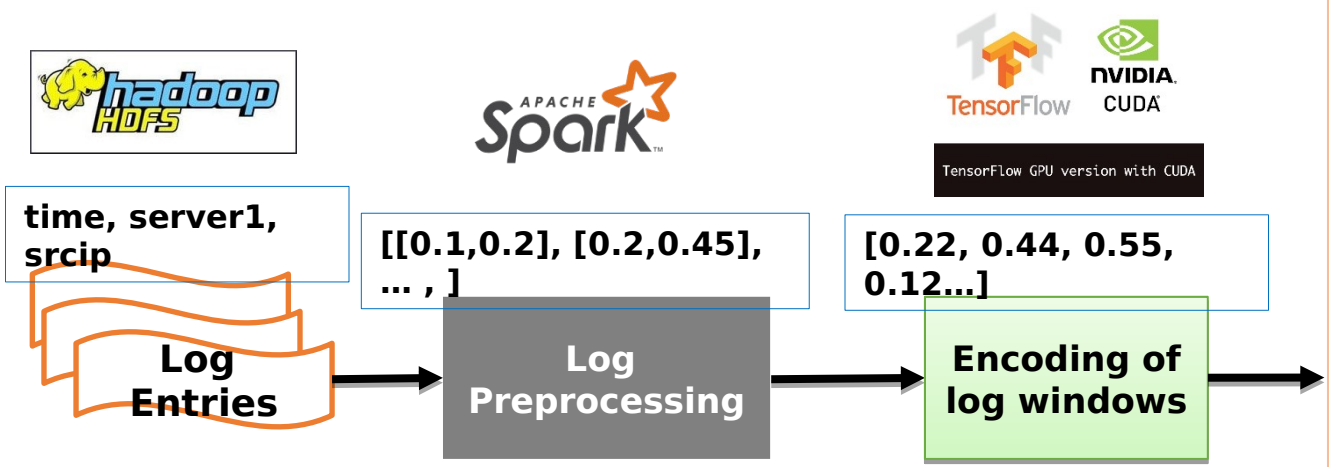
USING MACHINE LEARNING FOR INTRUSION DETECTION

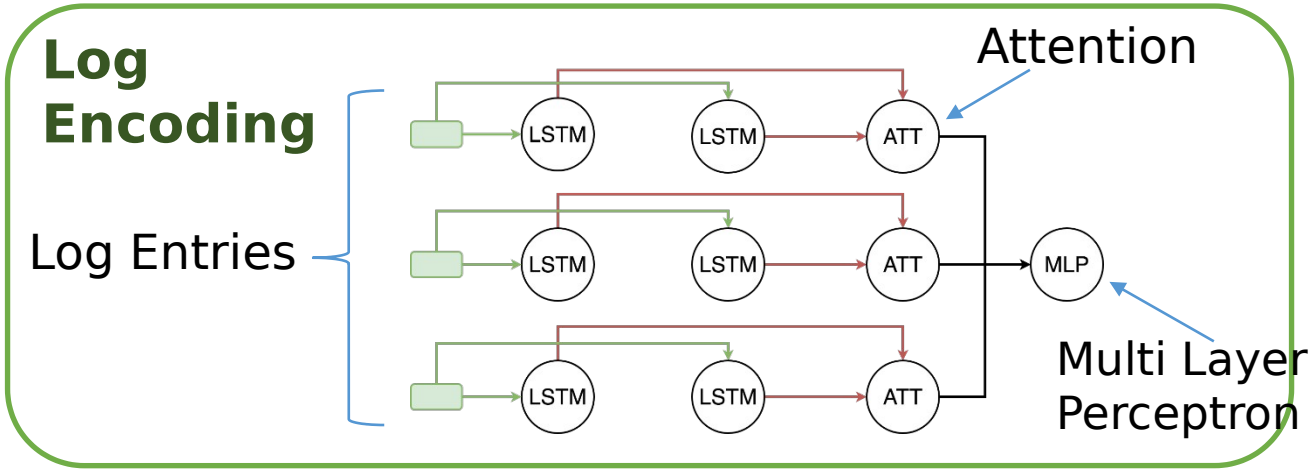
- Has the potential of detecting security incidents that can’t be easily detected using signature based techniques
- The model is trying to learn what is normal activity and detecting potential deviations from it

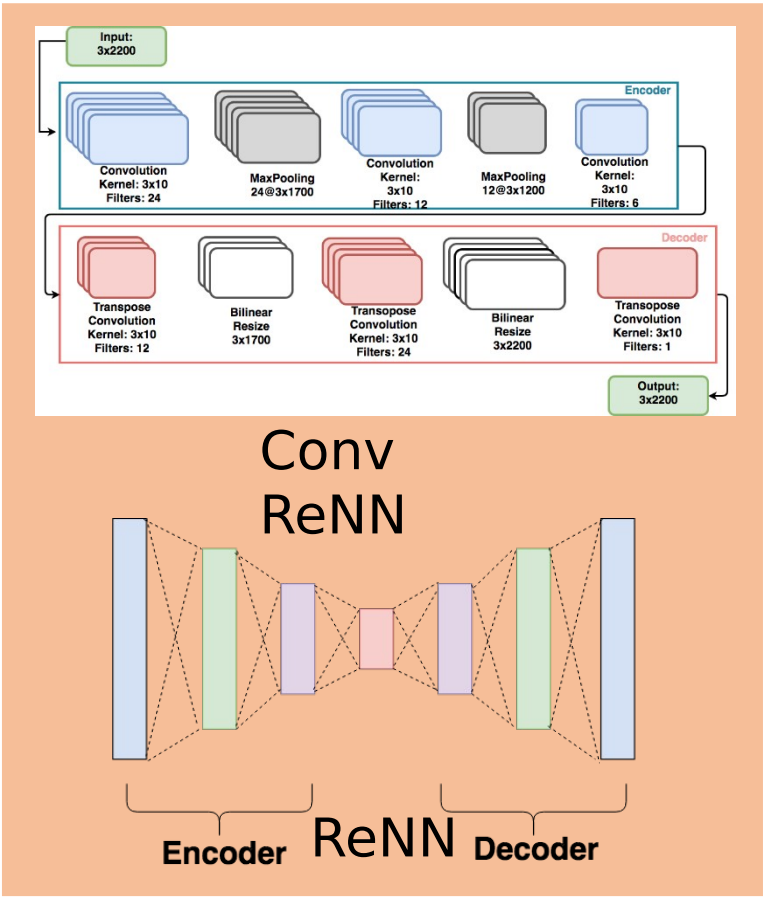
Challenges:

- No tagged data
- High rate of false positives
- Very challenging to define a baseline

MACHINE LEARNING PIPELINE







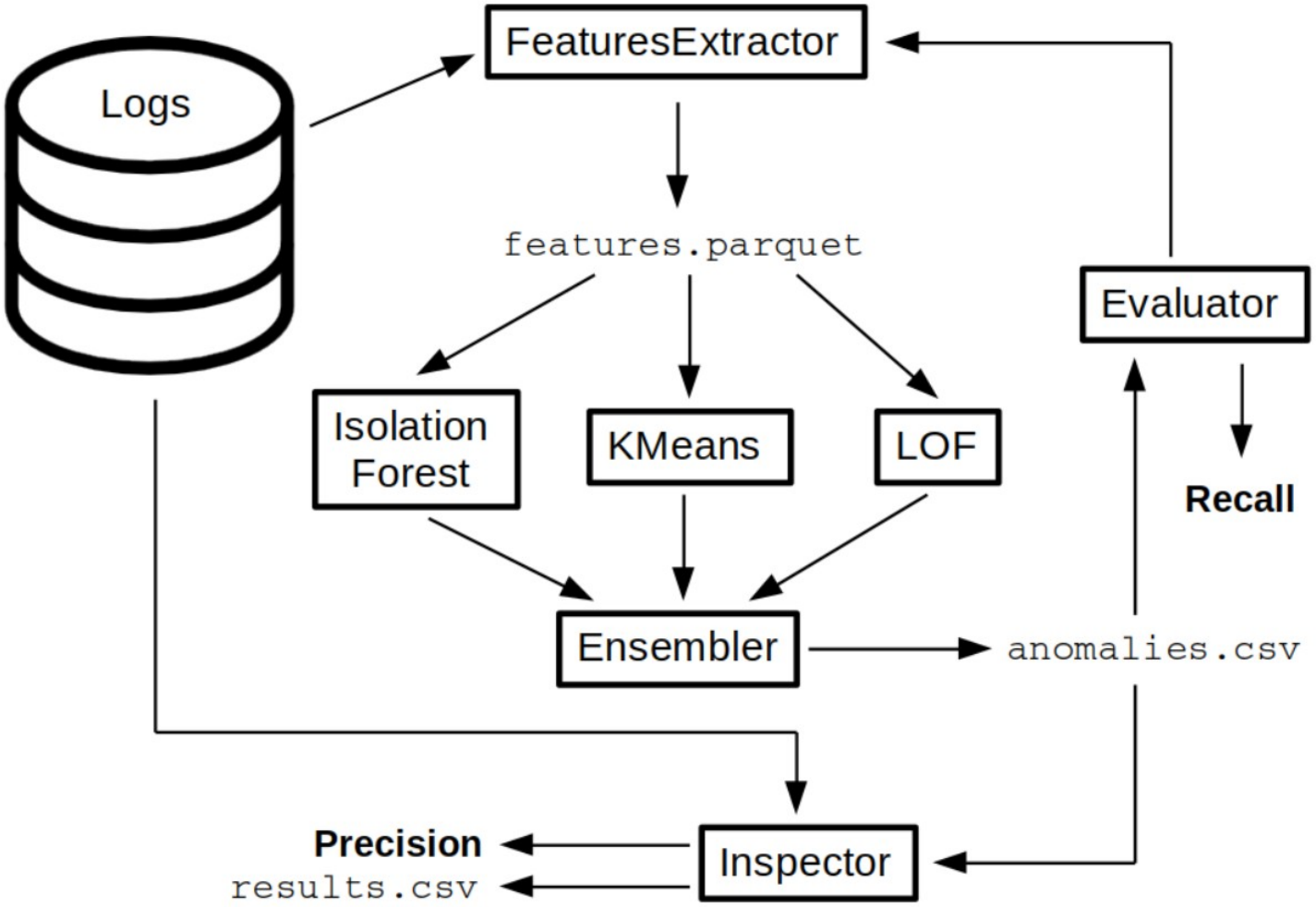
300 log entries with max reconstruction error

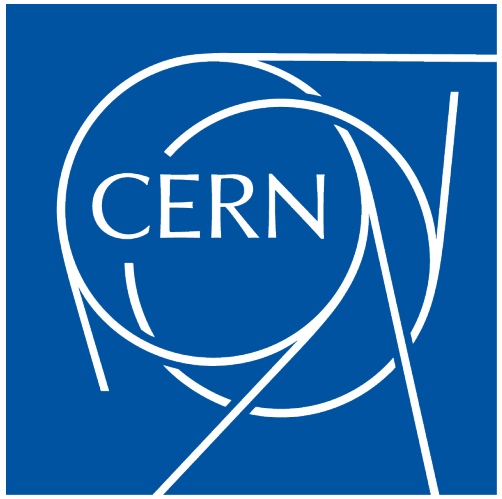
300 log entries with max reconstruction error

ANOMALY BASED INTRUSION DETECTION

- Uses Apache Spark, written in Scala
- Input from Apache Parquet files on HDFS
- 3 different anomaly detection algorithms being used:
  - Isolation Forest
  - K-means
  - Local Outlier Factor
- Recall and precision evaluation even without labelled test sets

ANOMALY BASED INTRUSION DETECTION




CERN