Draft contributions to the OpInt talk at next GDB

D. Bonacorsi (Univ. Bologna)
L. Decker De Sousa (Univ. Bologna)
F. Minarini (formerly Univ. Bologna)
L. Giommi (Univ. Bologna)
T. Diotalevi (Univ. Bologna)
+
B. Martelli (INFN-CNAF)
S. Rossi Tisbeni (INFN-CNAF)
F. Viola (INFN-CNAF)

Proposal 1

Add next (1) slide, probably around slide 21?

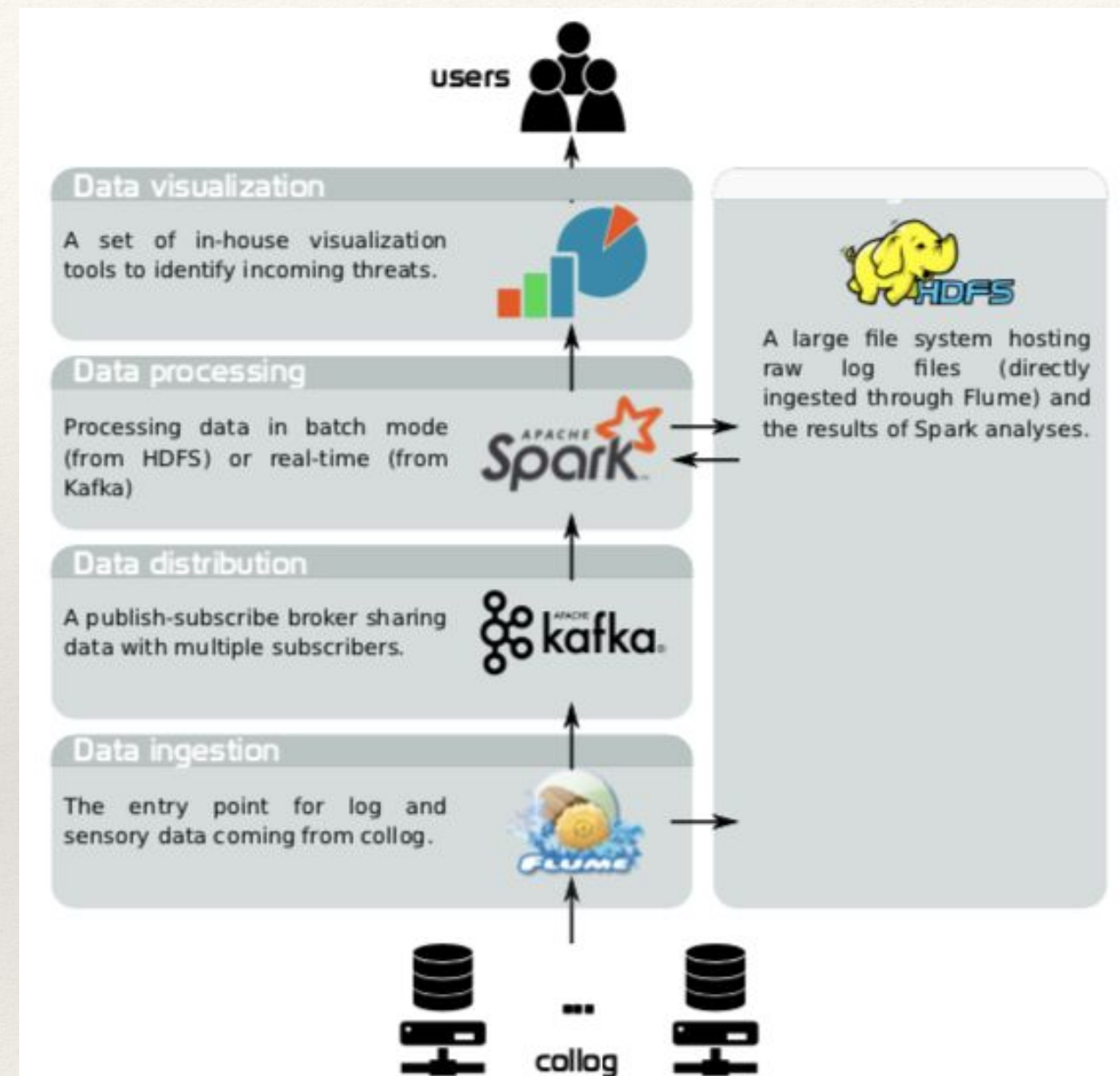(do we show infrastructure schemes of other sites?)

INFN-CNAF Monitoring and Analytics infrastructure

A novel **Big Data Analytics infrastructure** is being deployed at the INFN-CNAF Tier-1

- ✦ "Monitoring and Analytics at INFN Tier-1: the next step" (Martelli et al), presented at CHEP 2019 (Adelaide)

- ✦ "A big data infrastructure for predictive maintenance in large data centres" (Viola et al), presented at FRUCT25 2019 (Helsinki)

How to extract meaningful insight from these logs?

- ✦ i.e. around 40-50 M log entries produced per day by more than 1200 servers (with different logging strategies)

- ✦ work in progress by a pool of students (coord. Prof. Bonacorsi, University of Bologna) in exploring and comparing a variety of ML-enforced approaches before moving to prod

users

Data visualization
A set of in-house visualization tools to identify incoming threats.

Data processing
Processing data in batch mode (from HDFS) or real-time (from Kafka)

A large file system hosting raw log files (directly ingested through Flume) and the results of Spark analyses.

Data distribution
A publish-subscribe broker sharing data with multiple subscribers.

Data ingestion
The entry point for log and sensory data coming from collog.

collog

Proposal 2

Add next few slides around slide 21 of the draft talk - before or after the full-NLP approach

ML for log processing at a computing centre

Stimulated by the vast variety of logs from a Tier-1 (CNAF case), a small team of students at University of Bologna started to help CNAF personnel through quick "prototype and validate" cycles of a variety of ML-enforced techniques for log text processing and information extraction

✦ Not-exhaustive list of primary attempts:

- Supervised learning of labelled good/bad days in StoRM logs to predict future behaviour (L.Giommi et al)

- Collection and harmonization of system logs and prototypal Analytics services with the Elastic (ELK) suite (T. Diotalevi et al)

- Clusterisation of unstructured log entries based on measurements of Levenshtein distance (S. Rossi-Tisbeni et al)

- Unsupervised analysis and exploitation of volatility as a metric for an anomaly detection prototype in log-based predictive maintenance (F. Minarini et al)

- Log-agnostic template extraction on anomaly windows (L. Decker De Sousa et al)

Main goal: spot problematic services and raise warnings to operators before problems occur - without needing text processing (even less NLP) on the entirety of logs.

Other sites and/or interested individuals, are welcome to contact us and join!