Developing and Optimizing applications in Hadoop

Prasanth Kothuri, CERN IT Database Group

CHEP 2016
October 10-14
Outline

• What is Hadoop?
• Data Ingestion
• Data Formats
• Hadoop Processing Frameworks
  • Spark
• Batch / request-response application
• Troubleshooting
Hadoop

- A framework for large scale data processing
- **Distributed storage** and **distributed processing**
- Shared nothing architecture – scales horizontally
- Optimized for **high throughput on sequential data access**
Key aspects of Hadoop Application

- **Data Ingestion**: is the processing of bringing data to Hadoop ecosystem. We look at the configurations that deliver scalability, reliability and durability.

- **Data Formats**: has direct and high impact on the computations. We look at the criteria for choosing the right data format.

- **Processing Frameworks**: We look at the Apache Spark frameworks, its wide library support and parallel processing primitives both imperative and declarative.

- **Troubleshooting**: We present the tool developed to profile distributed applications.
Conclusion

• Data Ingestion, formats and processing framework are key aspects of building Hadoop Application

• Out of the myriad of Hadoop tools available, it is possible to build Hadoop Application using Kafka, Parquet and Spark

• hprofiler solves the challenge of identifying bottlenecks in distributed application

• ElasticSearch / Kibana can be leveraged to deliver User Interface
Discussion / Feedback

Q & A