

Comparison of Hadoop Framework Distributions

Petr Todorov, Supervisor: Ota Novotný, Faculty of Informatics and Statistics of University of Economics, Prague



Motivation

- The **Data** Produced by Both the Companies and the Individuals is **Growing Rapidly**
- **Real-Time Processing of Big Data Streams** is a Must
- Is There a **Hadoop Distribution Capable** of Continuous (Near) Real-Time Processing of Big Data Efficiently? If So, Which One to Use?

Goals

- **Analyze** the Apache Hadoop Framework Distributions Available
- Perform **Comparison** of Hadoop Distributions Regarding Real-Time Big Data Processing Possibilities

Big Data, Hadoop & Hadoop Distros

- Thorough Overview of **Big Data, Apache Hadoop** Including its Ecosystem, Hadoop **Distributions Market Evolution** and **Current State**

(Near) Real-Time Big Data Processing (Spark Structured Streaming)

- Real-Time ingestion:
 - **500 000** Tweets Mentioning **American Primary Election** Candidates Ingested During **2020 Super Tuesday**
 - **Sentiment Analysis** via AFINN
- Near Real-Time analysis:
 - **Overall** and **One-Hour Window Sentiment Statistics** Analyzed in Every Hadoop Distribution Compared (**CDH, MDP, HDP**)



Conclusion & Contribution

- Hadoop Distributions **Market Has Closed** -> No Deployments Without Subscription Any More
- **Real-Time Big Data Ingestion Proof of Concept** (Spark Structured Streaming)
- **No Clear Winner** in Hadoop Distributions Comparison -> Focus on Acquisition and Operational **Costs**

References

[1] KOMMENDA, N. 2020. Democratic primary delegate count – latest : Joe Biden and Bernie Sanders are competing to become the Democrats' nominee for president. *The Guardian* [online]. Guardian New & Media. [Cit. 14. 3. 2020]. Available from z: <https://www.theguardian.com/us-news/ng-interactive/2020/mar/18/democratic-primary-delegate-count-latest>.

Sentiment Analysis VS Real Results

