# Big-Data Analytics Architecture for Businesses: Open-source Perspective

Mert Onuralp Gökalp, Kerem Kayabay, Mohamed Zaki

Cambridge Service Alliance (CSA)
Institute for Manufacturing
University of Cambridge

January 2018

# Why study open-source tools for Big Data?

- Open-source tools have become the standard Big Data processing platforms*
- The gap: Study the open-source tools considering both managerial and technical perspective
- Some Questions
  - Why people prefer commercial solutions rather than open-source?
  - Do we need commercial Big Data solutions?

# The Big Tools Era

- Many tools continue to emerge to deal with big data at a fast pace
  - Characteristics: Volume, speed, diversity
  - Problems: Processing, storage, manipulation, aggregation, visualization
- Some tools only aim to analyse data in a certain domain
  - Internet of Things, Edge Computing
- Just by reviewing open-source tools, we have come across 6500 such tools and filtered down to 241

# Open-source big data tools

- Typically supported by companies that provide services over Internet
  - Google, Yahoo, Twitter, LinkedIn
- To provide better services to their users and third-party customers
- The tools are made available to IT industry as open-source tools
- They have become the standard big data processing platforms

# Some example tools

| Big Data Processing | Big Data Characteristic | Tools and Technologies |
|---|---|---|
| Batch processing | Volume | Hadoop, Spark, Flink |
| Stream processing | Velocity | Storm, Samza, S4, Spark Streaming, Flink Streaming |

| Big Data Storage | Big Data Characteristic | Tools and Technologies |
|---|---|---|
| NoSQL | Variety | MongoDB, Cassandra, Hbase, Redis |

# Choosing the right tool set

- Choices depend on the characteristics of data and domain of operation
- Businesses incur costs trying to adopt new technologies: technical debt
  - Training the workforce
  - Change existing source code to run on newer versions
  - Change the underlying toolset

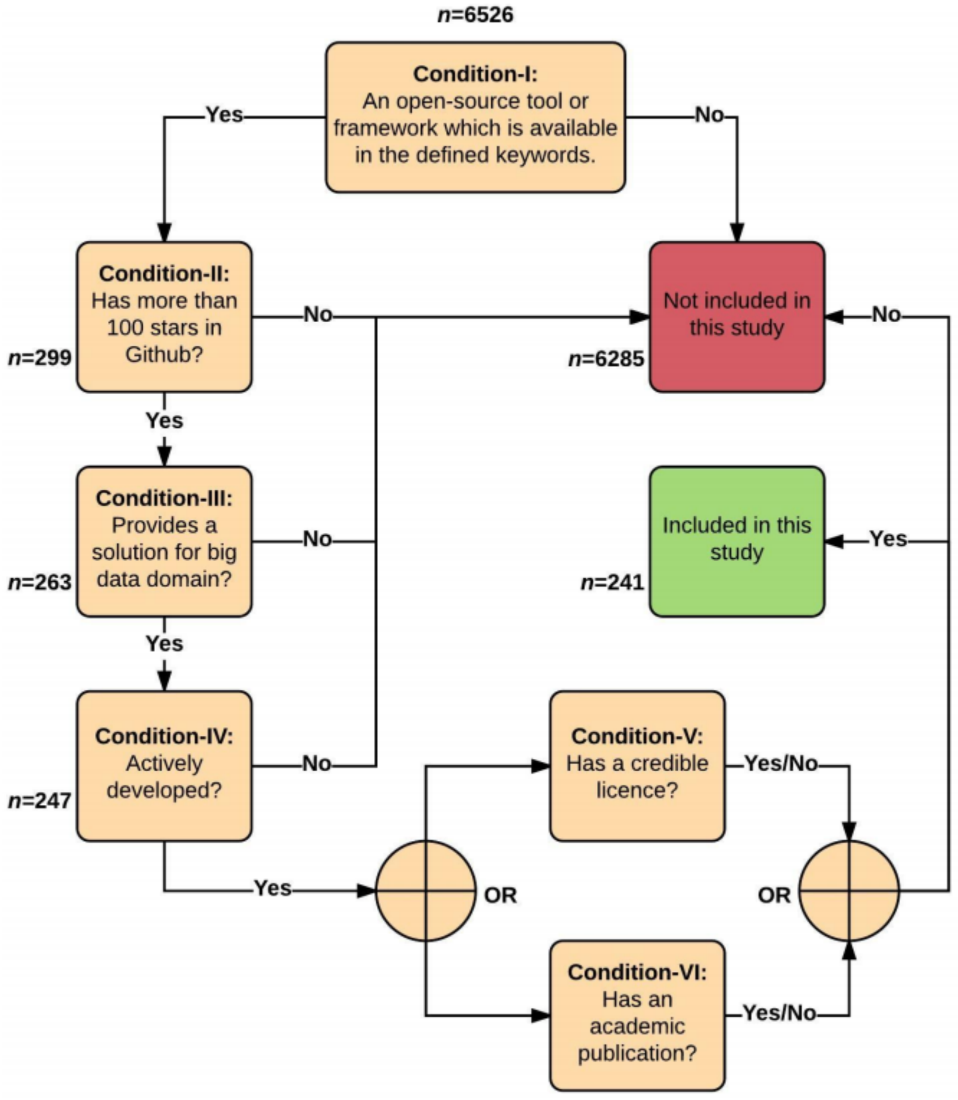
# Why firms outsource?

- Tracking the developments in this domain is hard
  - Most of the tools are unknown to the business world
- Not to lag behind the hype BUT commercial solution providers
  - Rely on a subset of available open-source tools
  - Do not have the domain specific expertise
  - Do not solve technical and soft challenges

# Aim of this study

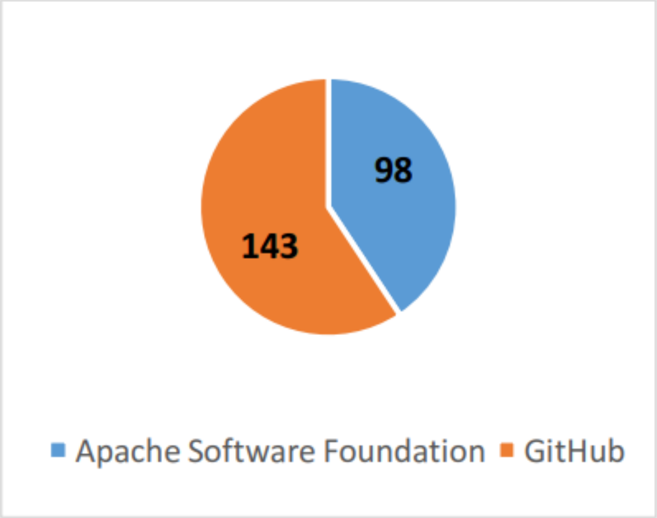
- Systematically review the open-source tools in the big data domain
- Establish a method for tracking the developments for the open-source tools
- Build a reference open-source big data analytics architecture
- Analyse firms to give directions to businesses using the proposed architecture

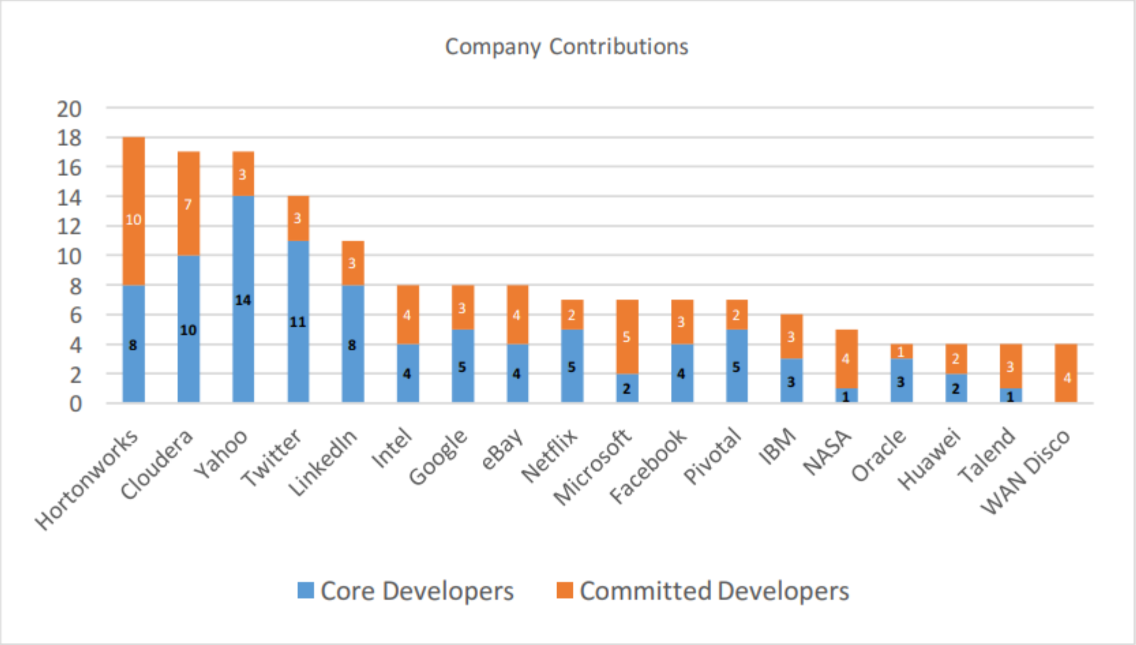
# Tool selection process



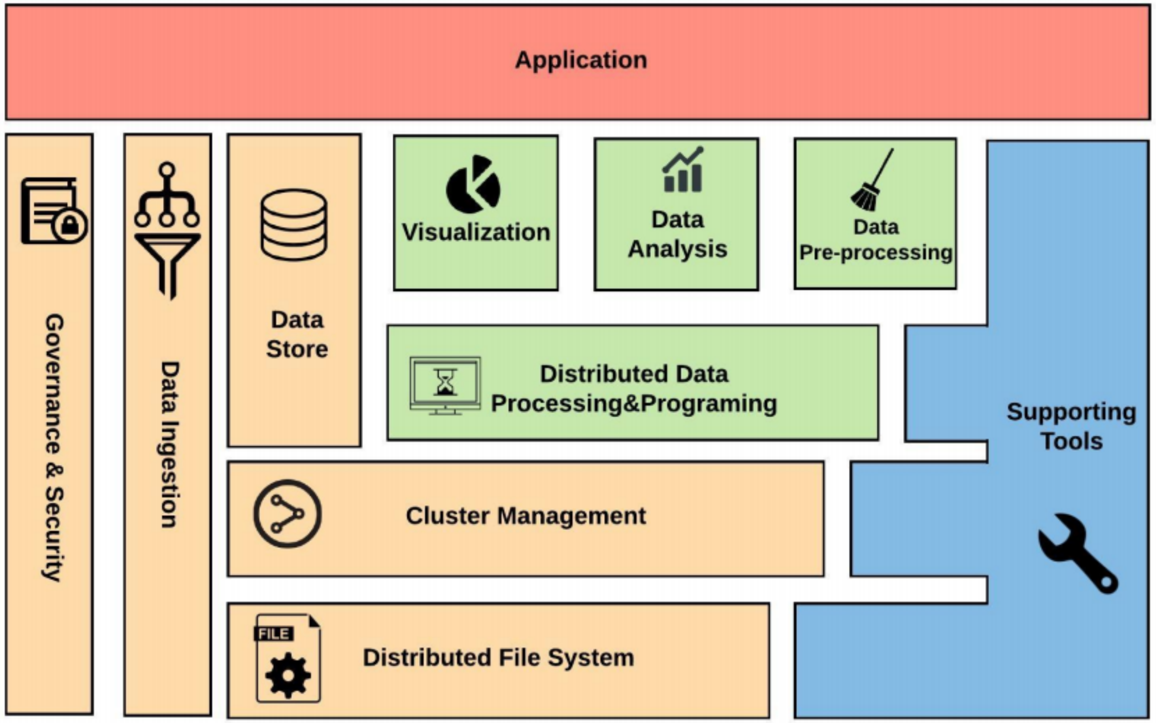
UNIVERSITY OF CAMBRIDGE
Cambridge Service Alliance

# Some figures

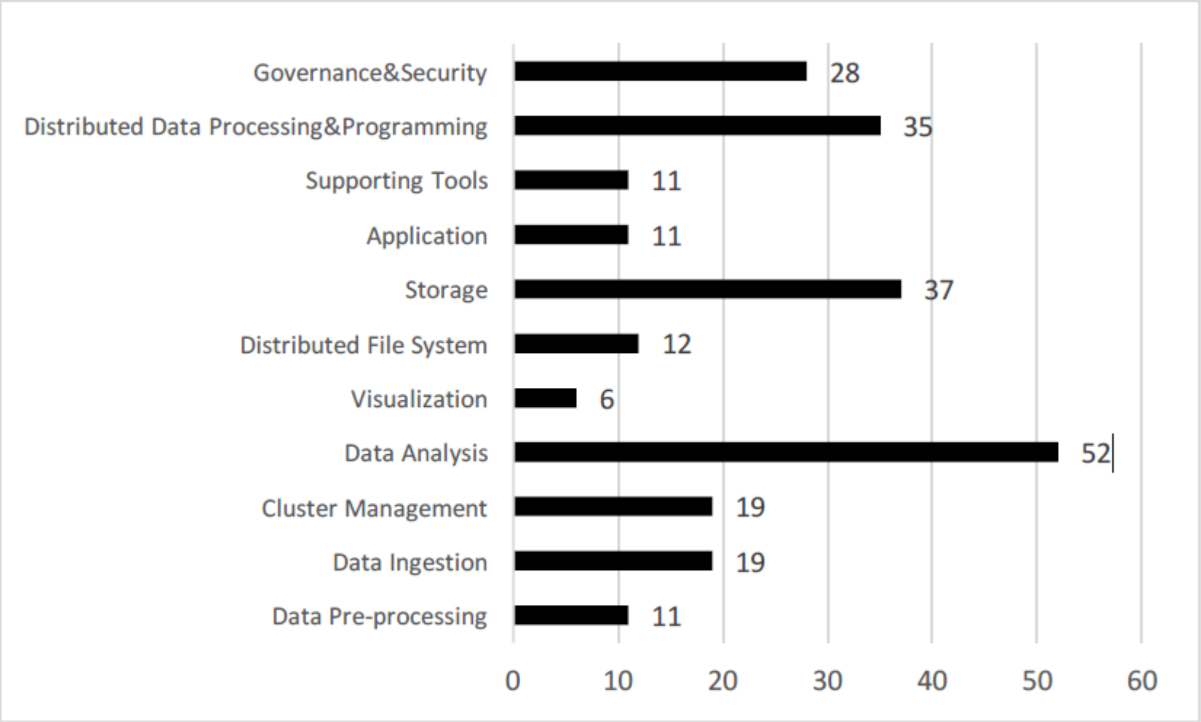




# Open-source Architecture



# Distribution of Open-Source Tools



UNIVERSITY OF CAMBRIDGE

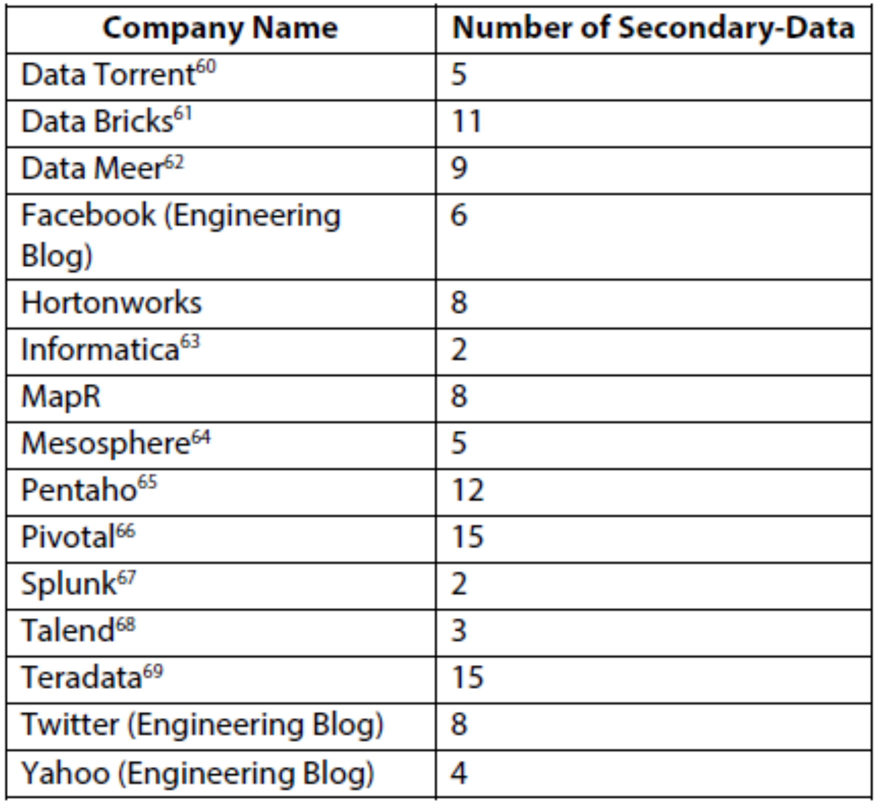
Cambridge Service Alliance

INFORMATICS INSTITUTE

# How to choose a big data tool?

- We need to come up with criteria
- We can look at (113) real-world use cases, solution briefs, whitepapers & blog posts from a wide range of industries
  - Telecommunication, healthcare, banking & finance, manufacturing, transportation, energy
- Secondary data-sets support the proposed big-data reference architecture

# Secondary use-case company distribution



| Company Name | Number of Secondary-Data |
|---|---|
| Data Torrent[60] | 5 |
| Data Bricks[61] | 11 |
| Data Meer[62] | 9 |
| Facebook (Engineering Blog) | 6 |
| Hortonworks | 8 |
| Informatica[63] | 2 |
| MapR | 8 |
| Mesosphere[64] | 5 |
| Pentaho[65] | 12 |
| Pivotal[66] | 15 |
| Splunk[67] | 2 |
| Talend[68] | 3 |
| Teradata[69] | 15 |
| Twitter (Engineering Blog) | 8 |
| Yahoo (Engineering Blog) | 4 |

# How to choose a big data tool?

- **Timing requirement:** Batch vs stream processing
- **Data size:** In-memory vs on-disk processing
- **Platform independency:** Interoperability of a big data tool
- **Data storage model:** Graph-based, key-value-based, document-based, time-series-based

# Problems of architecture development in big data

- Choosing the best tool
  - Abundance of tools
  - No single best tool
  - Maturity of a tool
- Domain-specific challenges
  - The gap between domain-specific knowledge and data science
- Firm-specific soft challenges

# Problems of architecture development in big data

- Firm-specific soft challenges
    - Managerial skills deep-rooted in an organization
    - Lack of data-driven organizational culture
    - Customers may not be able to perceive the value of big data

# To sum up

- Newer tools never cease to emerge in this domain
- We can foresee where the industry will focus research efforts
- Organizations should try to build their own big data architecture
    - Rely on open-source tools instead of imposed commercial solutions

# To sum up

- Organizations should try to build their own big data architecture
- It is rewarding
  - Capture domain-specific knowledge
  - The process would build a data-driven culture and develop the right managerial skills
  - Better decision-making

# Thank you!

- Q&A

| Date<br>14:30hr BST | Topic | Invited speaker |
|---|---|---|
| 2018 | | |
| Jan 15th | Big Data Analytics Architecture for Business | Mert/Kareem/Mohamed |
| **Feb 12th** | **Digital Business Transformation and Strategy: What do we know so far?** | **Mariam Helmy Ismail Abdelaal** |
| Mar 12th | Does buyers' dependence translate into financial performance? An empirical analysis of manufacturer-service provider relationships | Ornella Benedettini |

UNIVERSITY OF CAMBRIDGE
Cambridge Service Alliance

INFORMATICS INSTITUTE