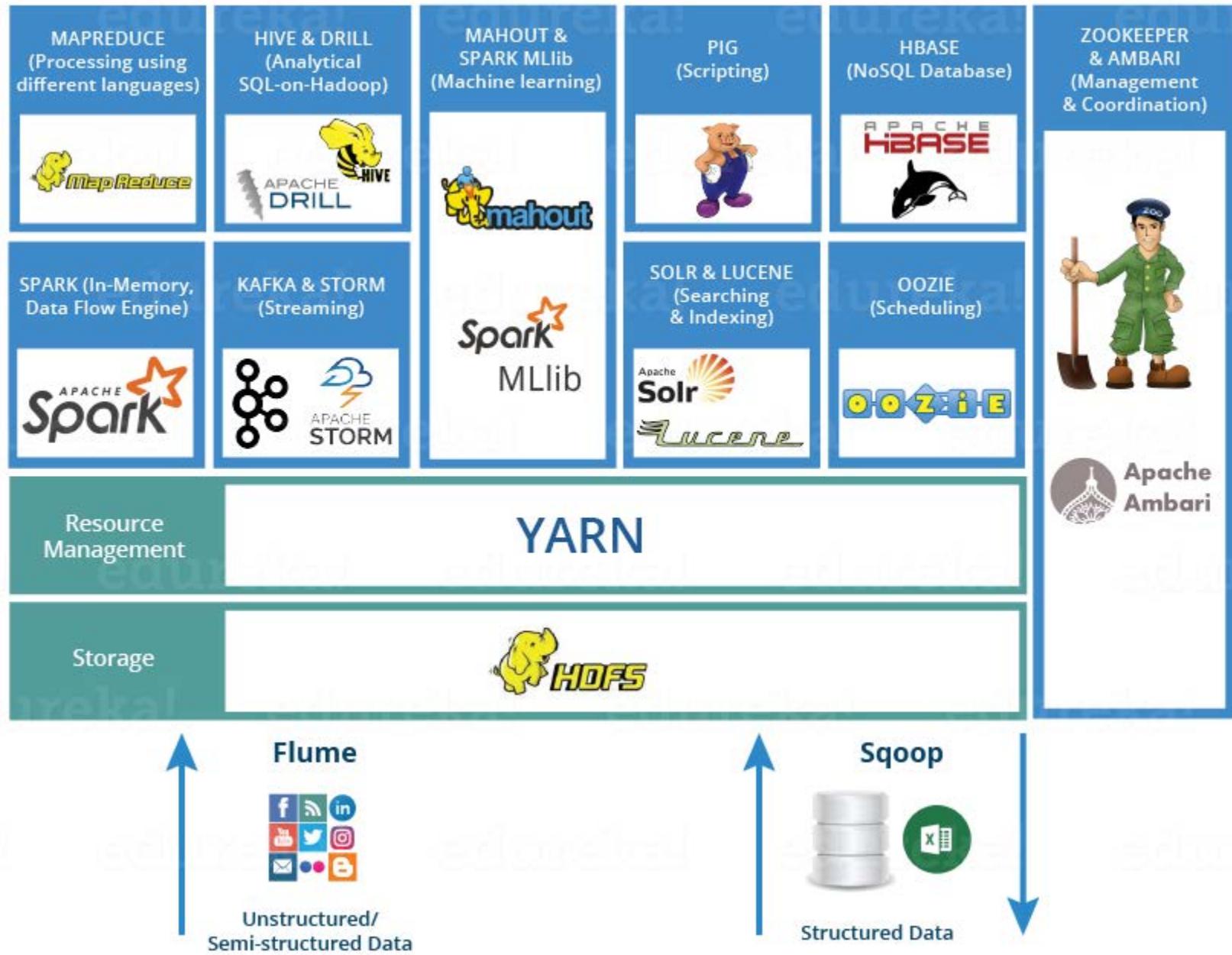




# Ekosistem Hadoop



# Ekosistem Hadoop

---

- Tools berhubungan dengan *ingest* (menelan/memasukkan) data
  - Unstructured/Semi-structure data
    - Flume
  - Structured Data
    - Sqoop
  - Streaming
    - Kafka
    - Storm
  - In-Memory, Data Flow Engine
    - Spark

# Ekosistem Hadoop

---

- Tools berhubungan dengan Penyimpanan
  - Pencarian dan Pembuatan Indeks (Seaching & Indexer)
    - Solr
    - Lucene
  - NoSQL Database
    - Hbase
    - MongoDB
  - Penyimpanan
    - Hadoop File System (HDFS)

# Ekosistem Hadoop

---

- Tools berhubungan dengan Pemrosesan
  - Pemrosesan (berbagai bahasa):
    - MapReduce
  - Scripting (untuk pemrosesan):
    - Pig
  - Analitik dengan SQL pada Hadoop:
    - Hive
    - Apache Drill
    - Apache Impala
  - Machine Learning:
    - Mahout
    - Spark MLlib

# Ekosistem Hadoop

---

- Tools berhubungan dengan *Management*
  - Resource Management
    - Yet Another Resource Negotiator (YARN)
  - Management & Coordination
    - Apache Ambari
    - Zookeper

# Ekosistem Hadoop

---

- Tools berhubungan dengan *Data Visualization*
  - Tableau
  - Zeppelin
  - D3.JS
  - Qlik
  - iCharts
  - R
  - SAS Visual Analytics
  - SAP Lumira

# Hadoop Distribution

---

- Meningkatnya penggunaan Hadoop beserta ekosistemnya menyebabkan persyaratan yang harus disesuaikan dengan kebutuhan suatu perusahaan/organisasi
- Akhirnya memunculkan **distribusi** komersial Hadoop
- **Distribusi** ini mengemas dukungan (*support*), fitur-fitur tambahan yang dirancang untuk **memudahkan penyebaran/instalasi** atau memungkinkan pengguna membuat analisis, keamanan, atau penanganan data tambahan ke dalam kerangka kerja.
- Beberapa vendor menyediakan solusi *on-site (on-premises)* atau *cloud-based*

# Hadoop Distribution

---

- Beberapa vendor (yang *on-site*) menyediakan *sandbox* dalam bentuk virtual machine (VM) supaya user bisa mencoba langsung tanpa melalui proses instalasi
- Namun, VM ini tidak seharusnya digunakan untuk *production*, jadi hanya mencoba dalam tahap *development*

# Hadoop Distribution

---

- Beberapa vendor Hadoop menyediakan solusi Big Data melalui distribusi masing-masing
  - Amazon **Elastic MapReduce**
  - **Cloudera** CDH Hadoop Distribution
  - **Hortonworks** Data Platform (HDP)
  - **MapR** Hadoop Distribution
  - IBM **Open Platform**
  - Microsoft **Azure's HDInsight** -Cloud based Hadoop Distribution
  - Pivotal Big Data Suite
  - **Datameer Professional**
  - **Datastax Enterprise Analytics**

# Hadoop Distribution

- Beberapa perbandingan antara distribusi Hortonworks, Cloudera, MapR (sumber: [experfy.com](http://experfy.com))

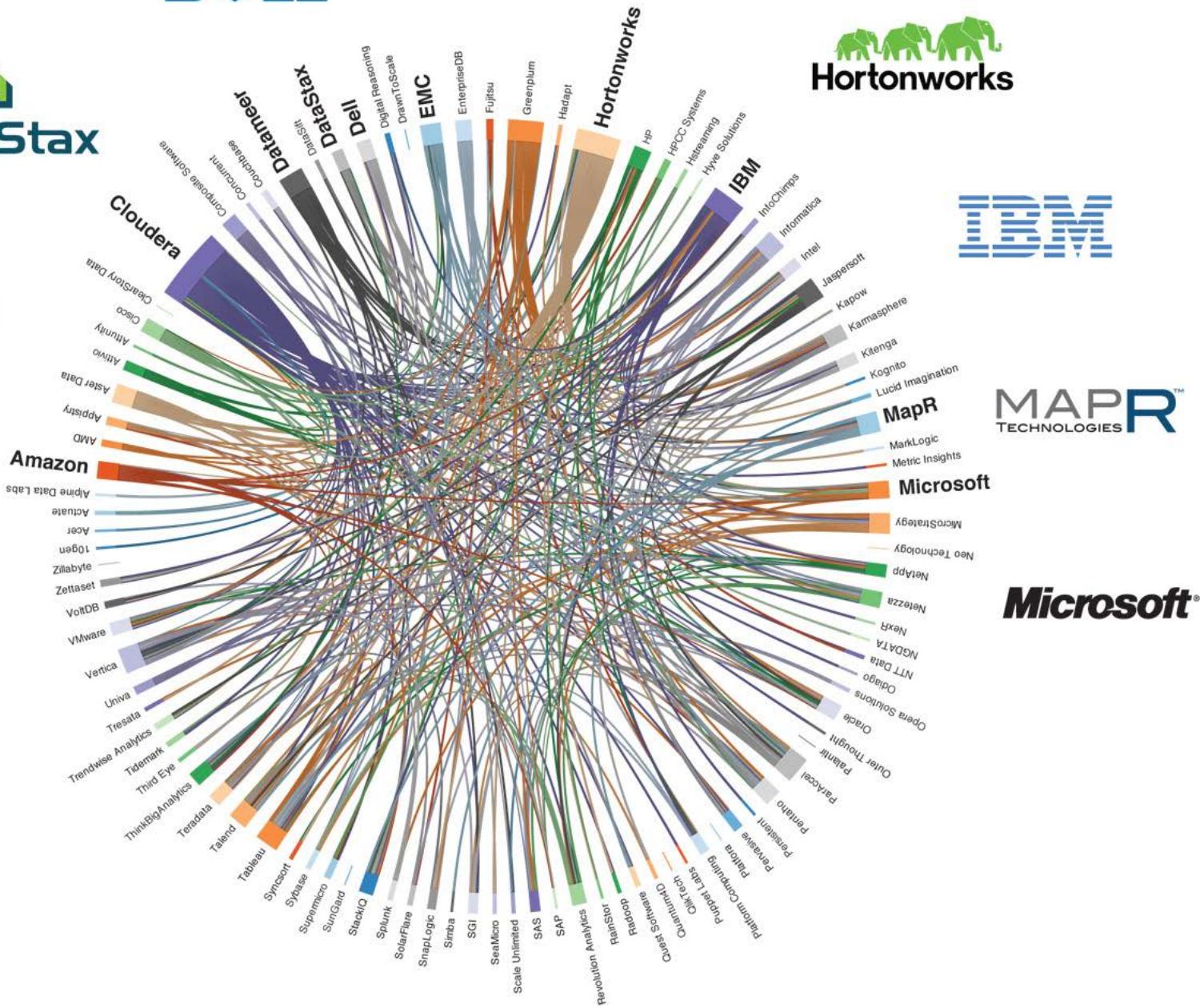
	Hortonworks	Cloudera	MapR
<b>Performance and Scalability</b>			
Data Ingest	Batch	Batch	Batch and streaming writes
Metadata Architecture	Centralized	Centralized	Distributed
HBase Performance	Latency spikes	Latency spikes	Consistent low latency
NoSQL Applications	Mainly batch applications	Mainly batch applications	Batch and online/real-time applications
<b>Dependability</b>			
High Availability	Single failure recovery	Single failure recovery	Self healing across multiple failures
MapReduce HA	Restart jobs	Restart jobs	Continuous without restart
Upgrading	Planned downtime	Rolling upgrades	Rolling upgrades
Replication	Data	Data	Data + metadata
Snapshots	Consistent only for closed files	Consistent only for closed files	Point-in-time consistency for all files and tables
Disaster Recovery	No	File copy scheduling (BDR)	Mirroring
<b>Manageability</b>			
Management Tools	Ambari	Cloudera Manager	MapR Control System
Volume Support	No	No	Yes
Heat map, Alarms, Alerts	Yes	Yes	Yes
Integration with REST API	Yes	Yes	Yes
Data and Job Placement Control	No	No	Yes
<b>Data Access</b>			
File System Access	HDFS, read-only NFS	HDFS, read-only NFS	HDFS, read/write NFS (POSIX)
File I/O	Append only	Append only	Read/write
Security: ACLs	Yes	Yes	Yes
Wire-level Authentication	Kerberos	Kerberos	Kerberos, Native



**EMC<sup>2</sup>**  
where information lives



cloudera



# Acknowledgment

---

## 1. Analisis Big Data: Putra Pandu A