

Armazenamento e Processamento de dados massivos



Armazenamento e Processamento de dados massivos

Cláudio Lúcio do Val Lopes

Armazenamento e Processamento de dados massivos

1

O que é exatamente isto?
Do que estamos falando ?

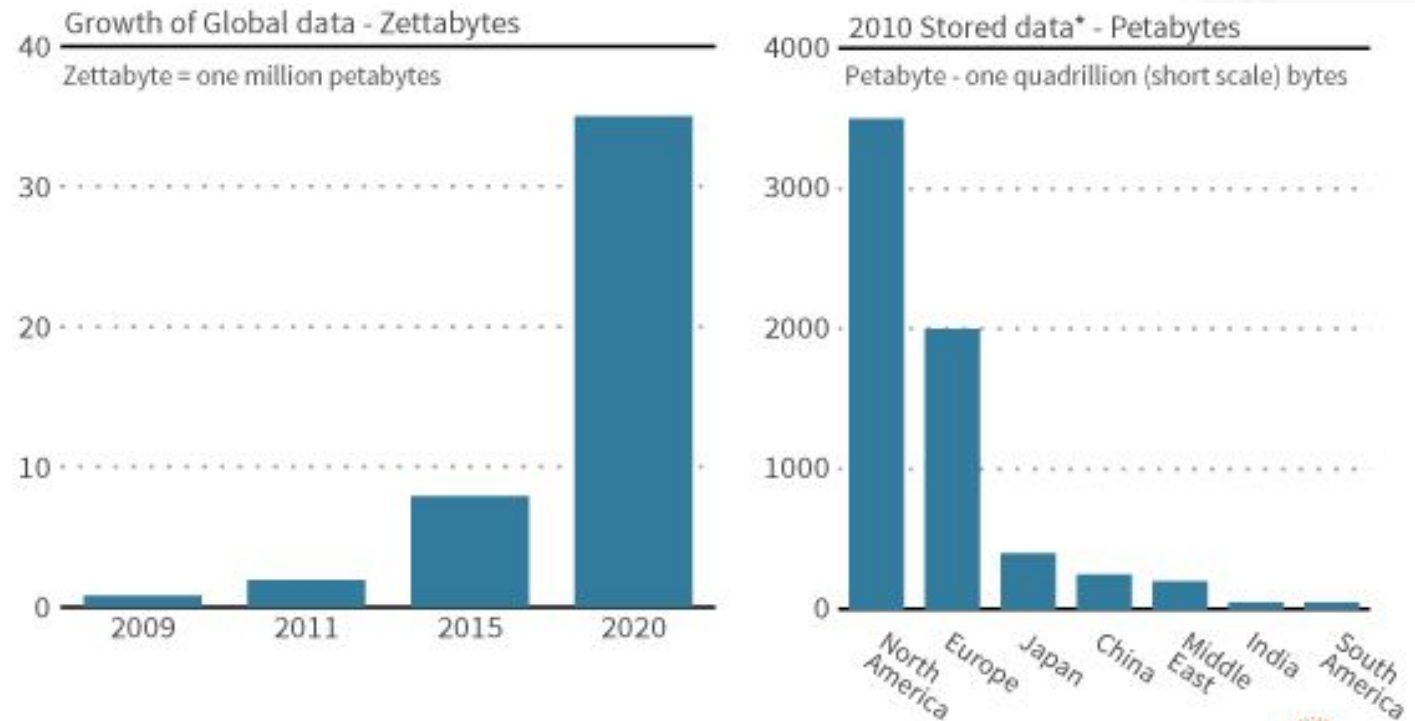
2

Algumas propostas atuais
Quais opções existem na prática?

3

Hands-on DataBricks
Aprenda na prática

O que é exatamente isto?



GPT-3

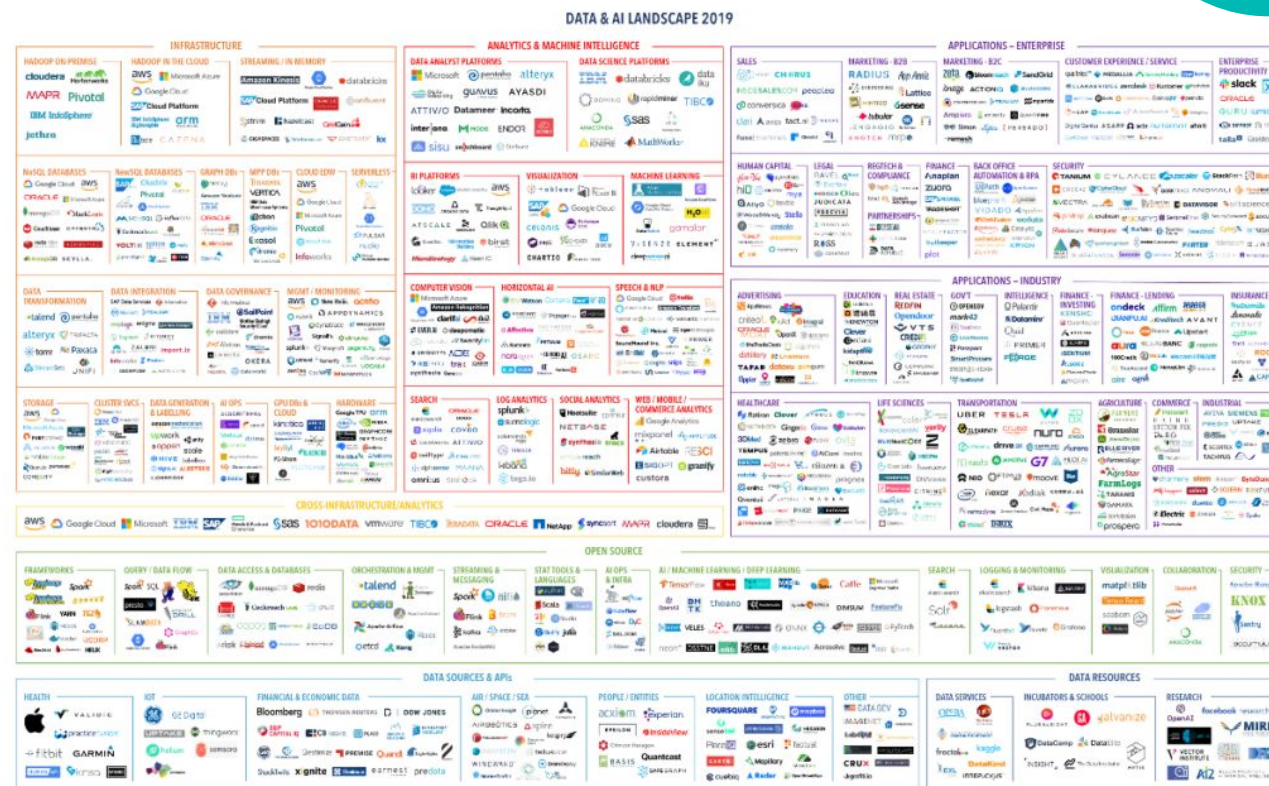
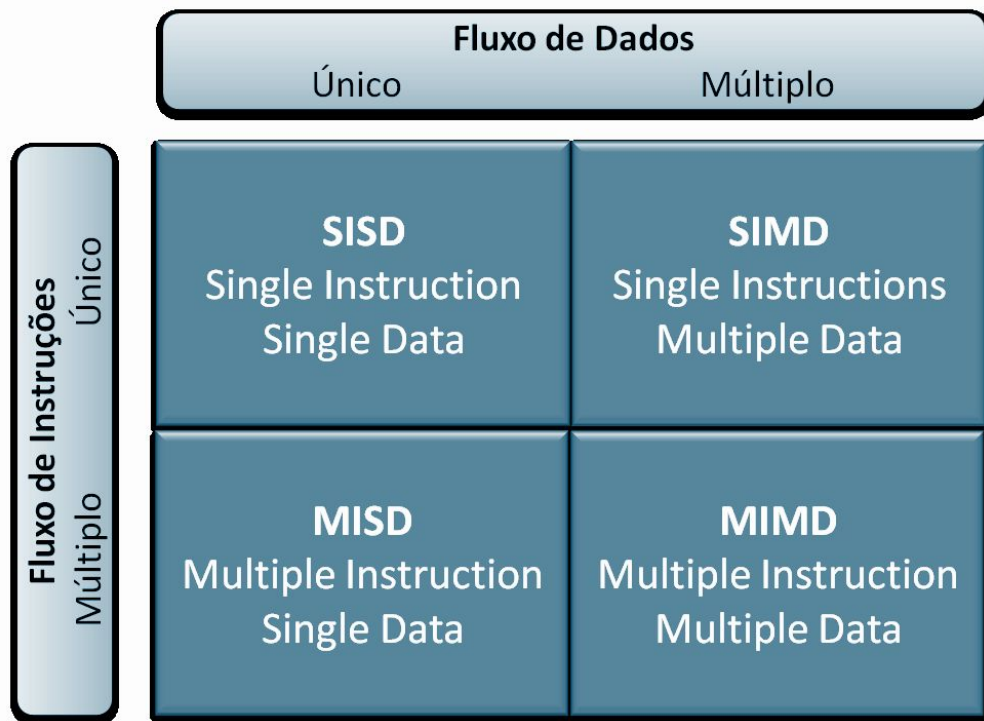
499 Bilhões de tokens

~ \$ 5 milhões para treinamento

175 Bilhões de parâmetros

Do que estamos falando?

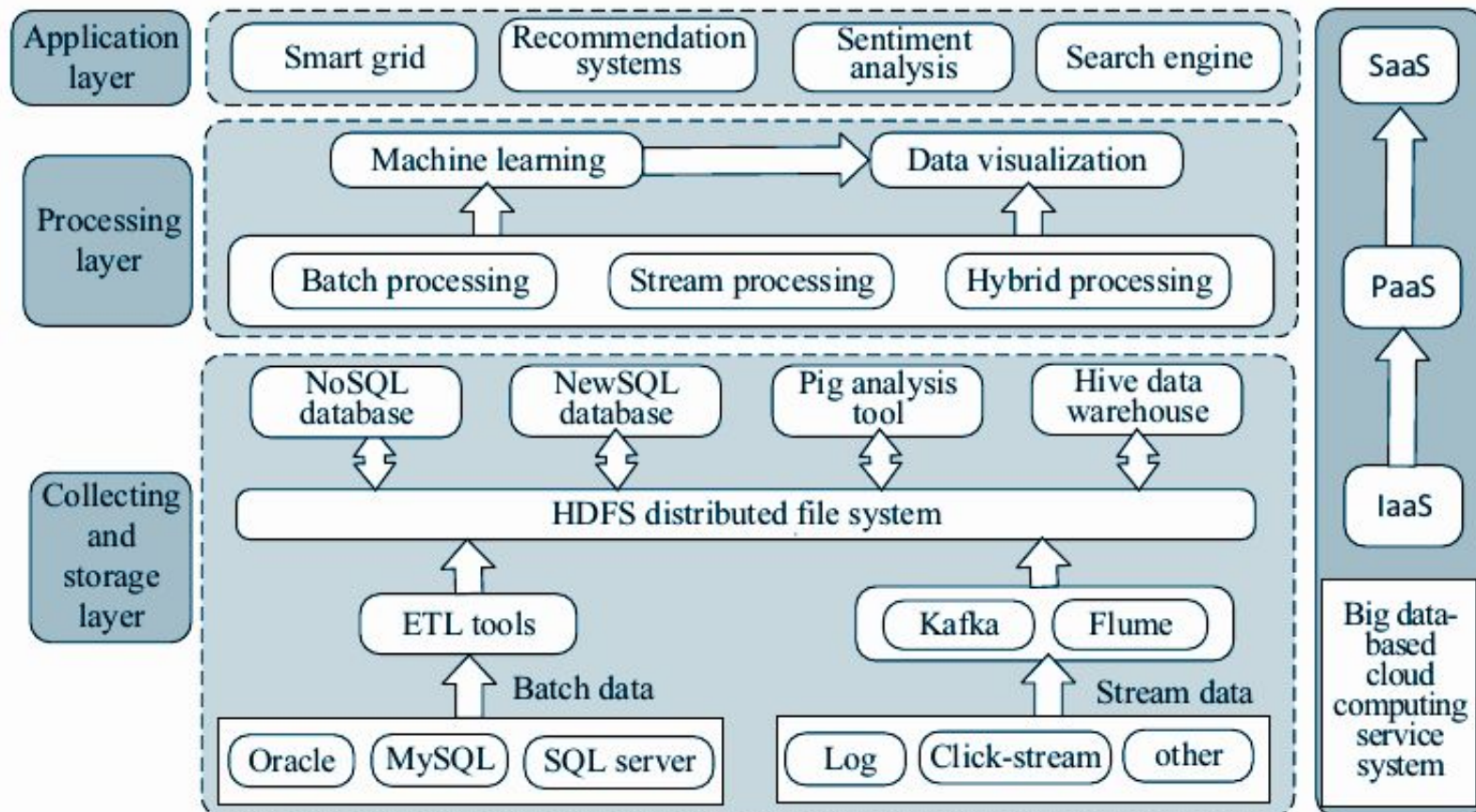
Taxonomia de Flynn



Propostas Atuais



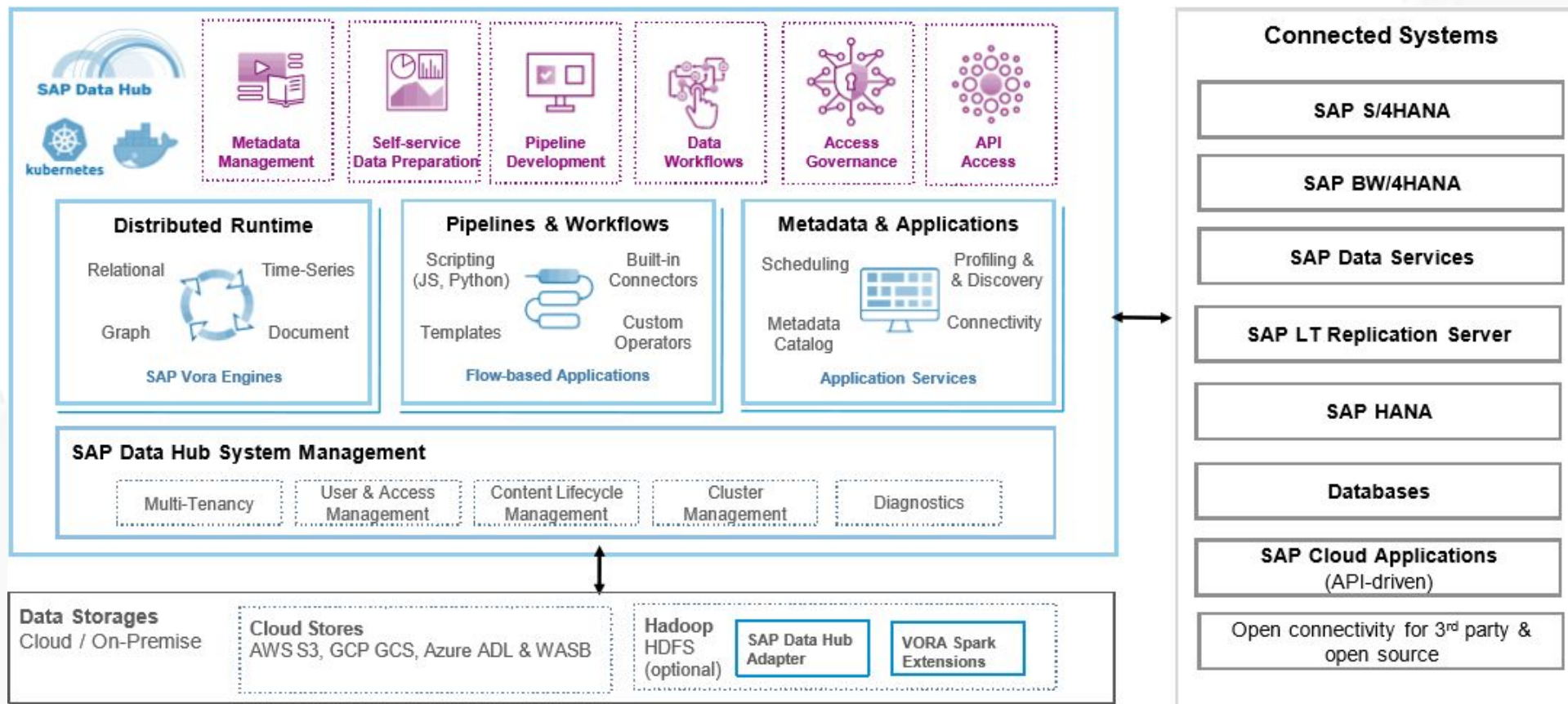
Propostas Atuais - Macro esquema



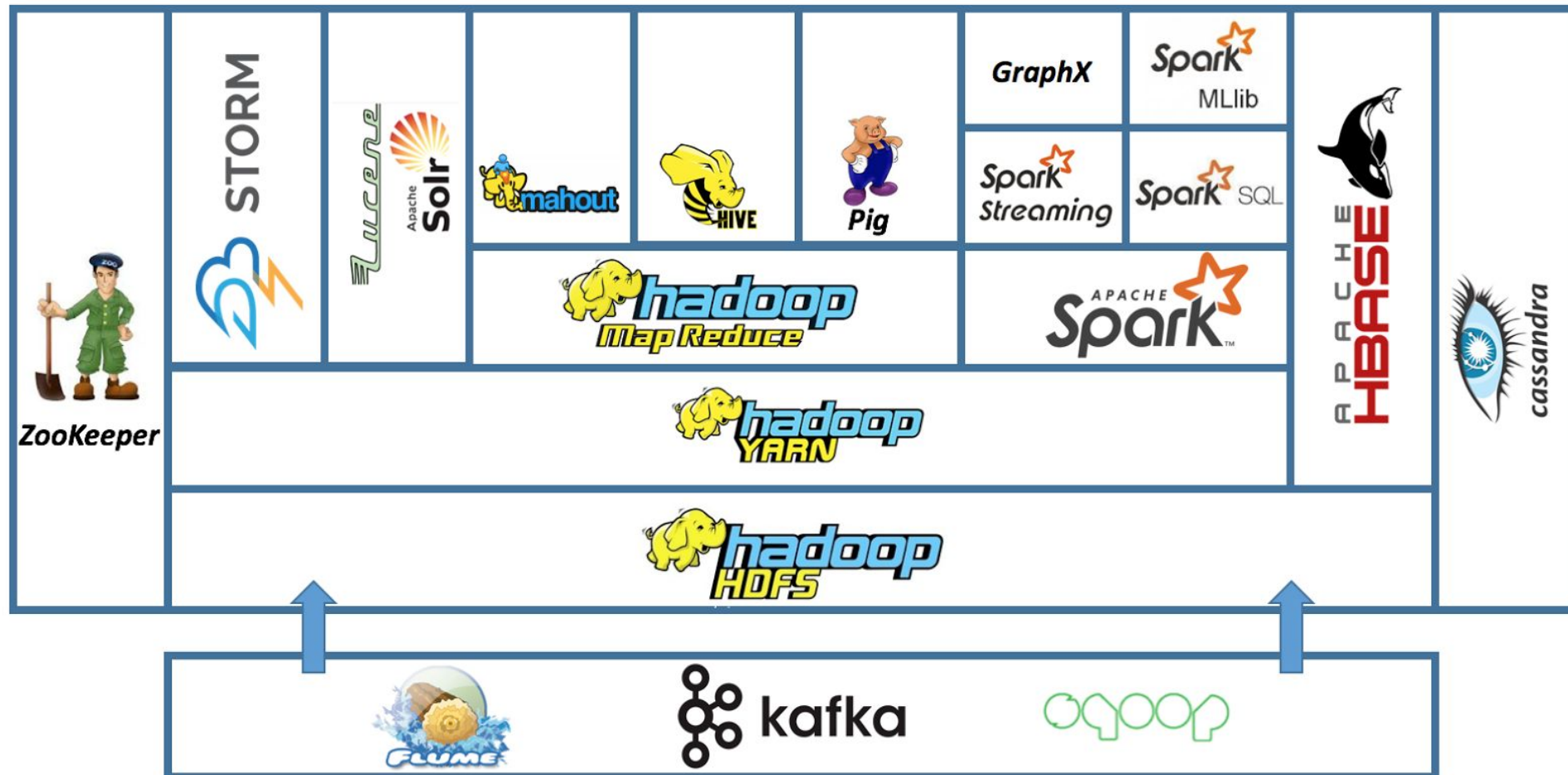
Opções na Prática - Open source

NoSQL databases	Types	Characteristics	Advantages	Disadvantages	Scenarios
Redis [21]	Key-Value	Key-value pairs; Simple data structure; High scalability; Low query update efficiency	Excellent read-write performance; Data persistence; Read-write separation	Without automatic fault tolerance and recovery function; Not supporting for on-line expansion	User information storage related to ID (key), i.e. sessions, configuration files, parameters
Memcached [22]			Memory-based data processing; Distributed and scalable mode; Fast processing speed	Small single cache data capacity; Not supporting for data persistence	
MongoDB [23]	Document-Oriented	Document-based data processing unit; Multiple data formats, i.e. JSON, XML	Fast access speed; Multiple data types; High query efficiency	Low read-write efficiency; Large space occupancy	Suitable for storing log files and analyzing data in real time
CouchDB [24]			Availability and concurrency; High flexibility	Low query efficiency; High maintenance difficulty	
Hbase [25]	Column-Family	Column-based data storage model; High data correlation in the same column family	Fast loading speed; Suitable for storage of large data; Efficient compression rate	Low multi-conditional query efficiency; Not directly supporting SQL statement query	Suitable for high concurrency operation on big data; Real-time read/write access
Cassandra [26]			Elastic scalability; Flexible schema; Multi-list data structures	Not supporting for ACID transactions and atomic operations	
Neo4j [27]	Graph-Oriented	Graph-based data storage formats, i.e. entities are vertices; High data correlation between entities; Efficient query performance	Efficient query performance; High performance graphics algorithm	Not supporting for large graph partition	Suitable for relational data storage, recommendation engines, community websites
GraphDB [28]			ACID transactions and atomic operations; Multi-model objects	High complexity	

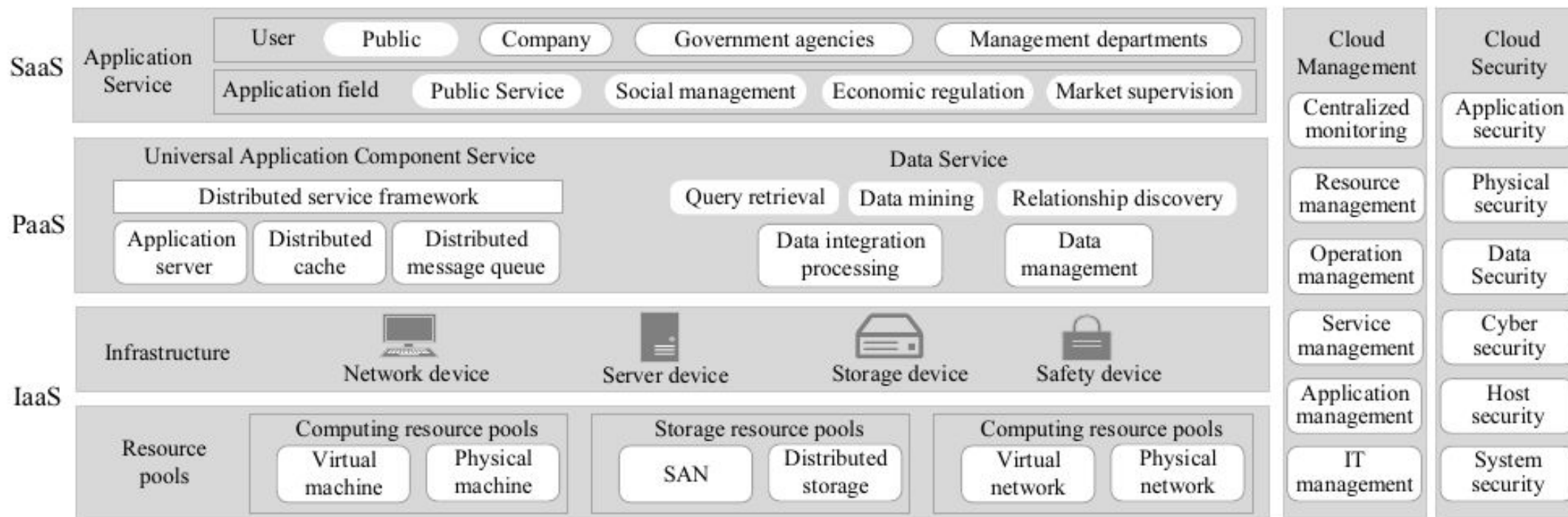
Opções na Prática - SAP



Opções na Prática - Hadoop



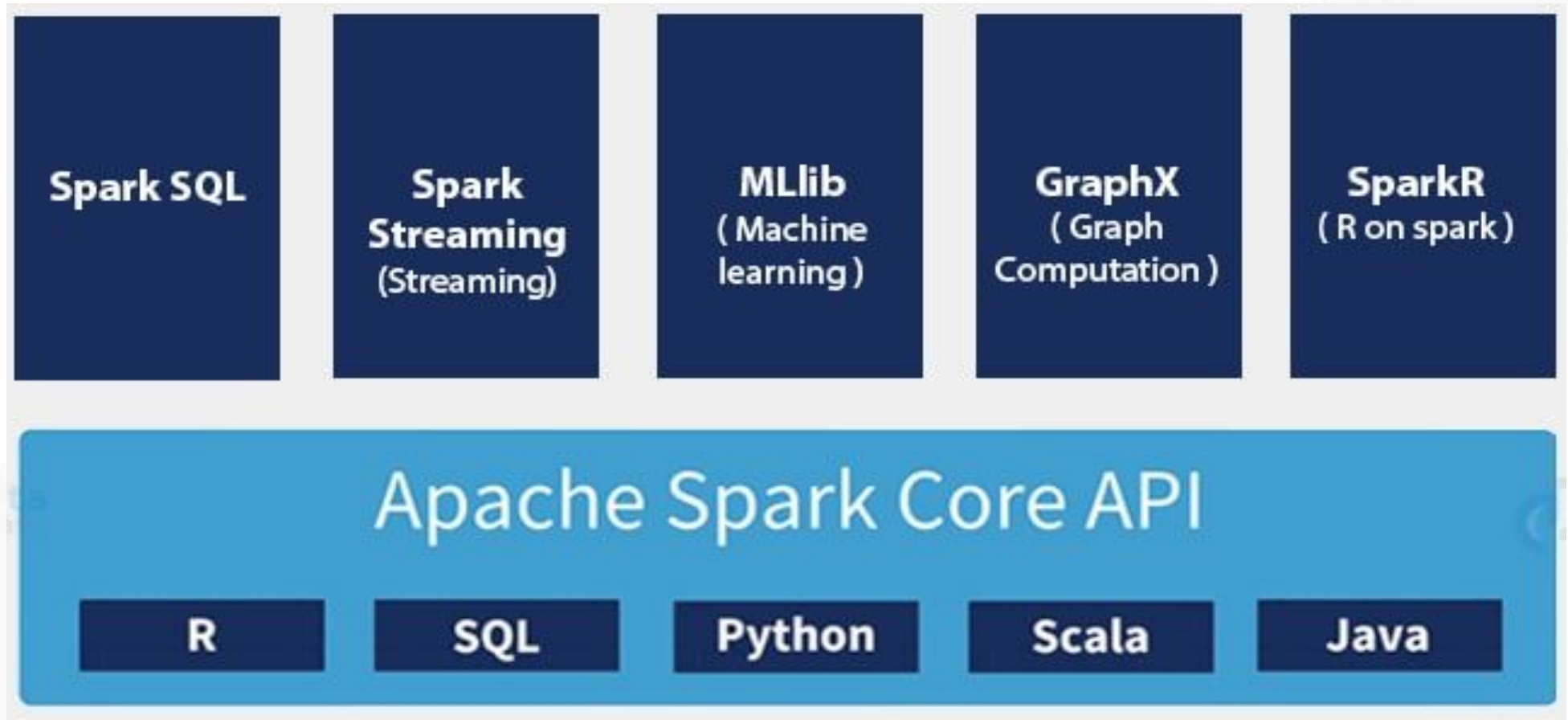
Opções na Prática - Cloud computing



Hands-on DataBricks



Hands-on DataBricks





Obrigado !!!

claudio.lucio@a3data.com.br

