Додаток А

Слайди презентації

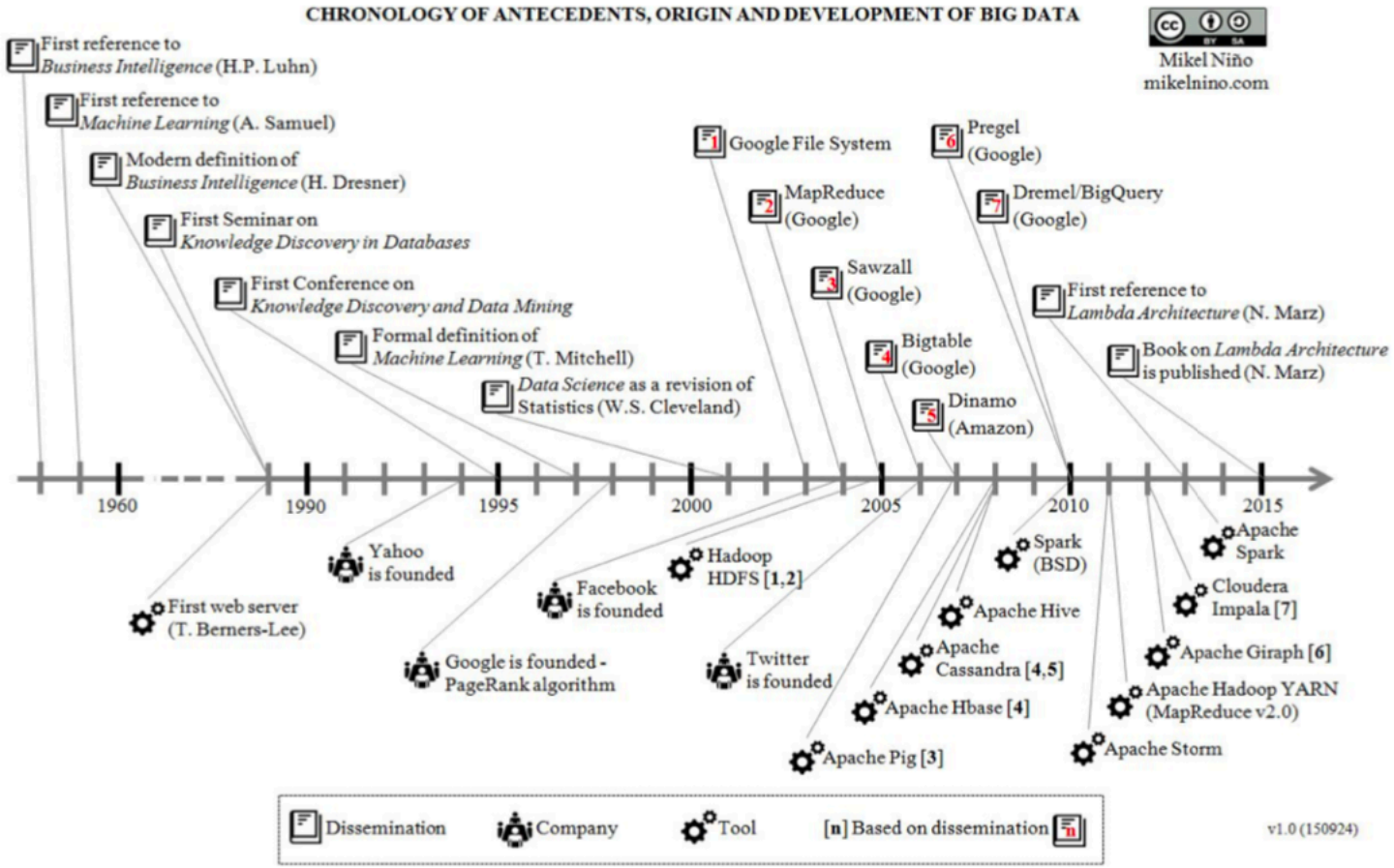
Дослідження методів обробки потоків даних у Big Data

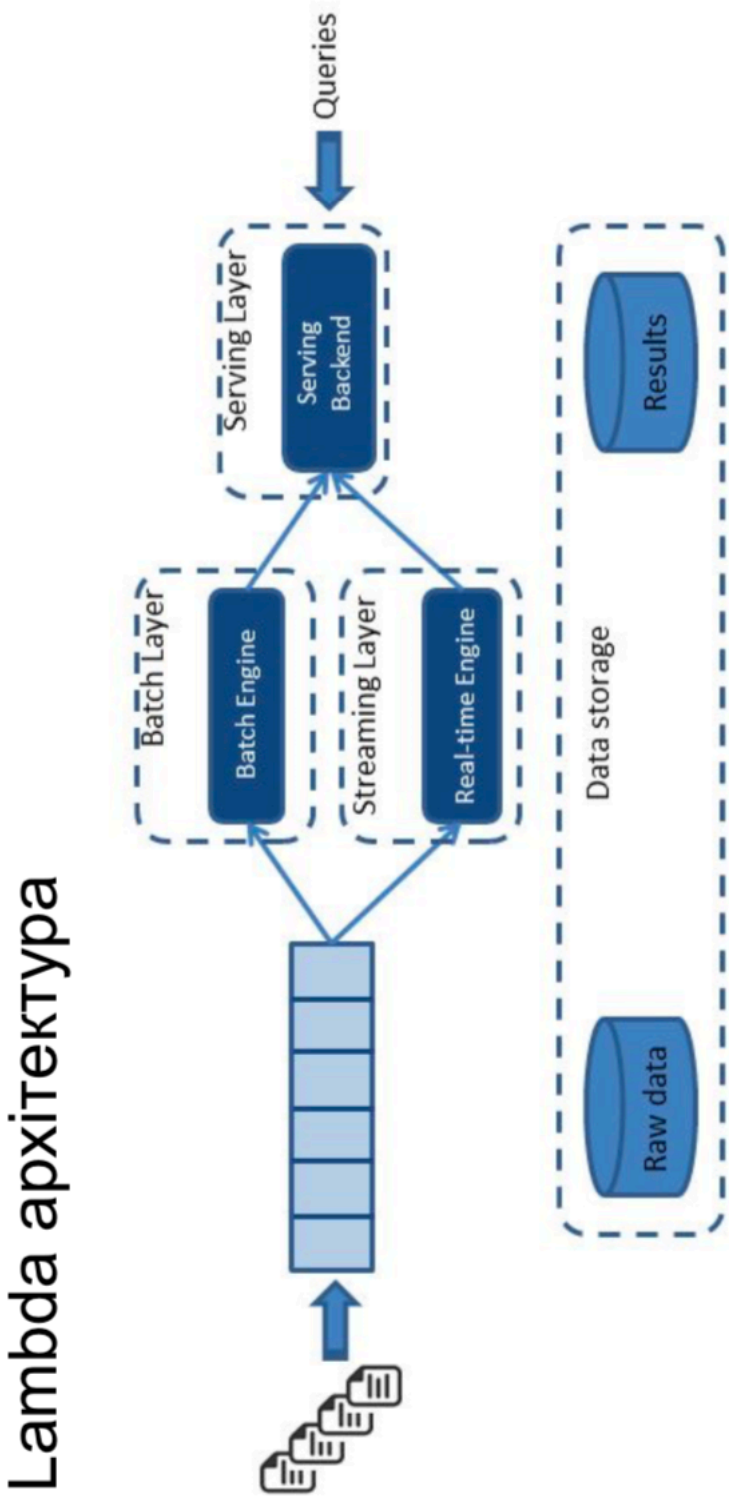
виконав
ст. гр. ІПЗм-17-1 Рукавиця А.С.
керівник
проф. каф ПІ Руткас А.Г.

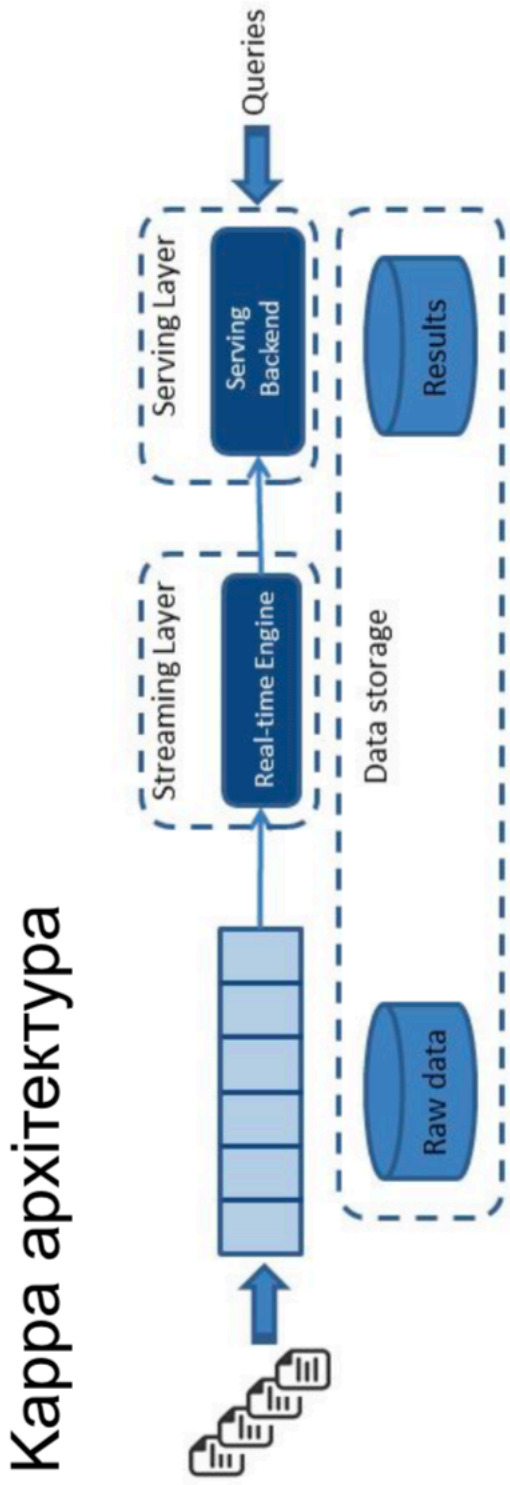
Мета роботи. Об’єкт і предмет дослідження

- Дослідження алгоритмів обробки потоків даних у режимі близькому до реального часу в хмарному середовищі із застосуванням паралельних та розподілених обчислень
- Дослідження способів забезпечення відмовостійкості при обробці потоків даних
- Дослідження способів обробки даних що приходять із запізненням
- Дослідження Big Data екосистем Spark, Flink та Kafka

Вплив Big Data

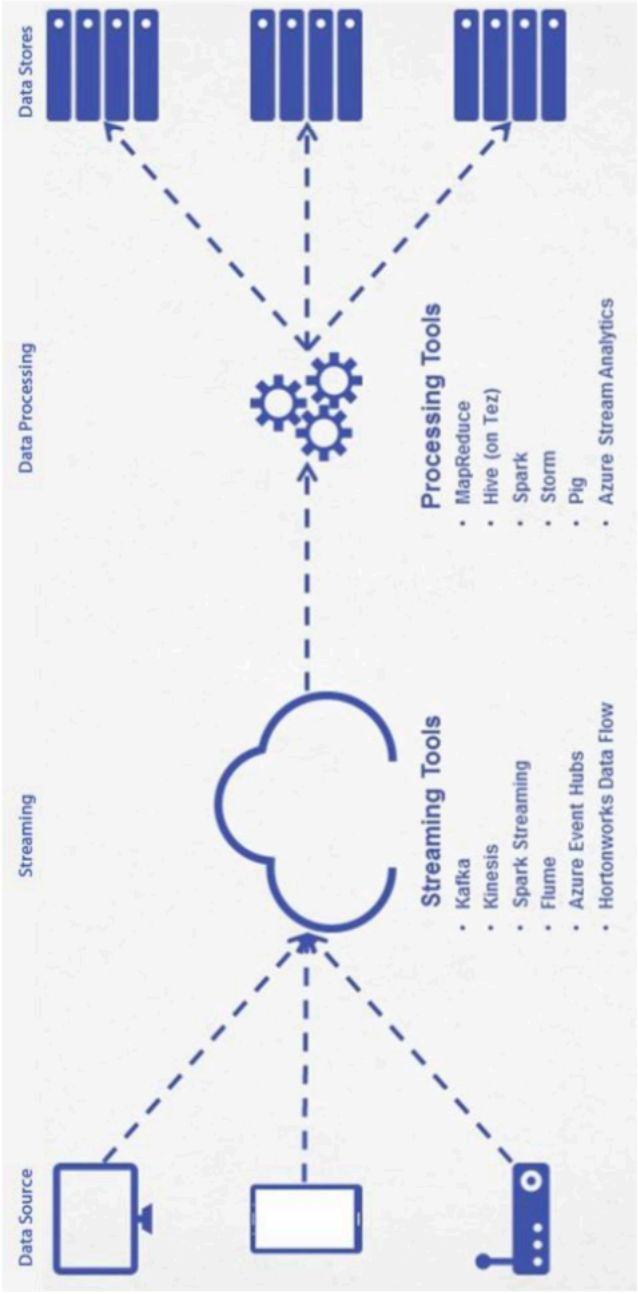

CHRONOLOGY OF ANTECEDENTS, ORIGIN AND DEVELOPMENT OF BIG DATA
Mikel Niño
mikelnino.com
First reference to Business Intelligence (H.P. Luhn)
First reference to Machine Learning (A. Samuel)
Modern definition of Business Intelligence (H. Dresner)
First Seminar on Knowledge Discovery in Databases
First Conference on Knowledge Discovery and Data Mining
Formal definition of Machine Learning (T. Mitchell)
Data Science as a revision of Statistics (W.S. Cleveland)
Google File System
MapReduce (Google)
Sawzall (Google)
Bigtable (Google)
Dinamo (Amazon)
Pregel (Google)
Dremel/BigQuery (Google)
First reference to Lambda Architecture (N. Marz)
Book on Lambda Architecture is published (N. Marz)
1960
1990
1995
2000
2005
2010
2015
First web server (T. Berners-Lee)
Yahoo is founded
Google is founded - PageRank algorithm
Facebook is founded
Hadoop HDFS [1,2]
Twitter is founded
Apache Pig [3]
Apache Hbase [4]
Apache Cassandra [4,5]
Apache Hive
Spark (BSD)
Apache Storm
Apache Hadoop YARN (MapReduce v2.0)
Apache Giraph [6]
Cloudera Impala [7]
Apache Spark
Dissemination
Company
Tool
[n] Based on dissemination
v1.0 (150924)


Lambda архітектура
Queries
Serving Layer
Serving Backend
Batch Layer
Batch Engine
Streaming Layer
Real-time Engine
Data storage
Results
Raw data

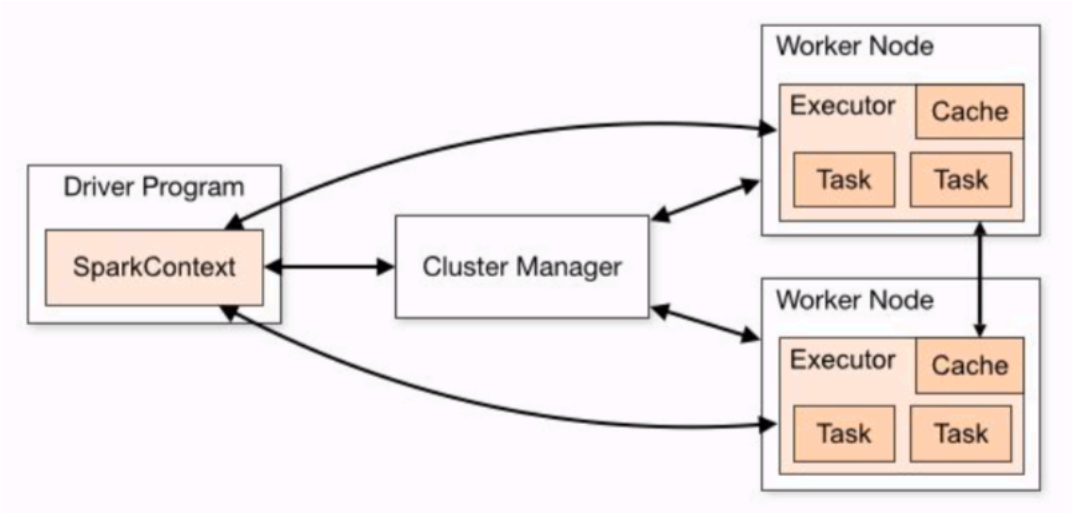


Каппа архітектура

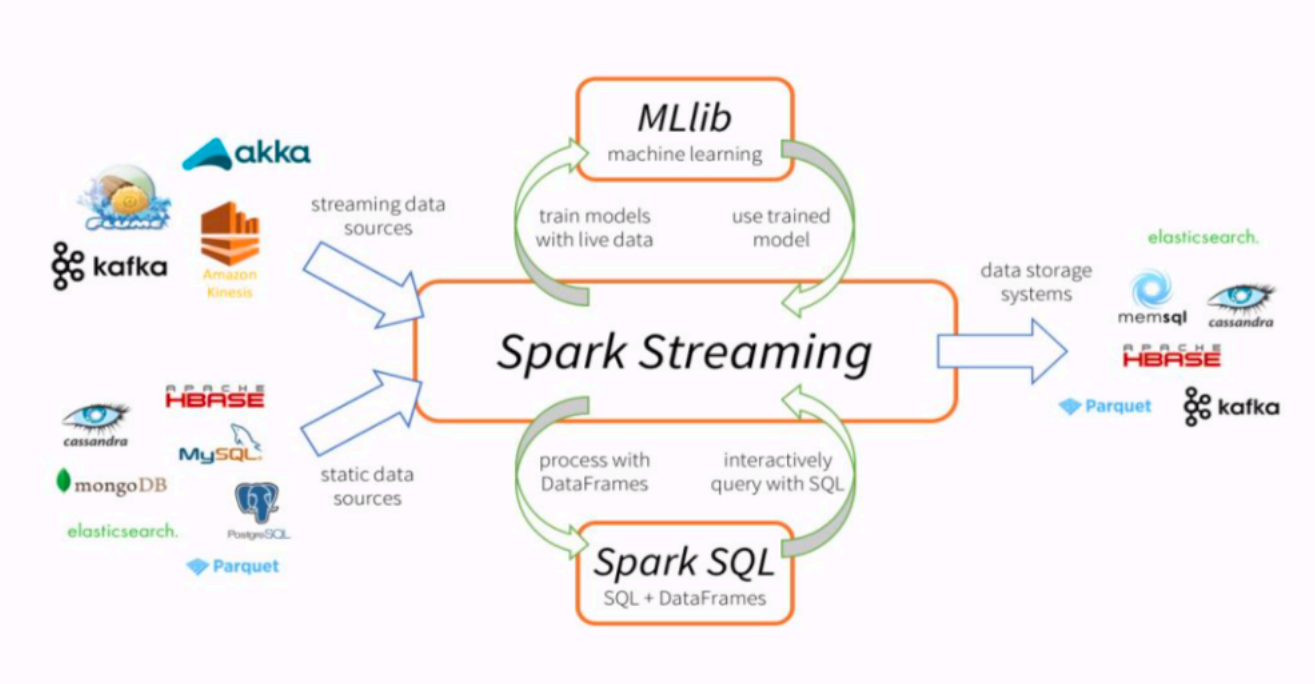
Потоки даних



Spark Cluster

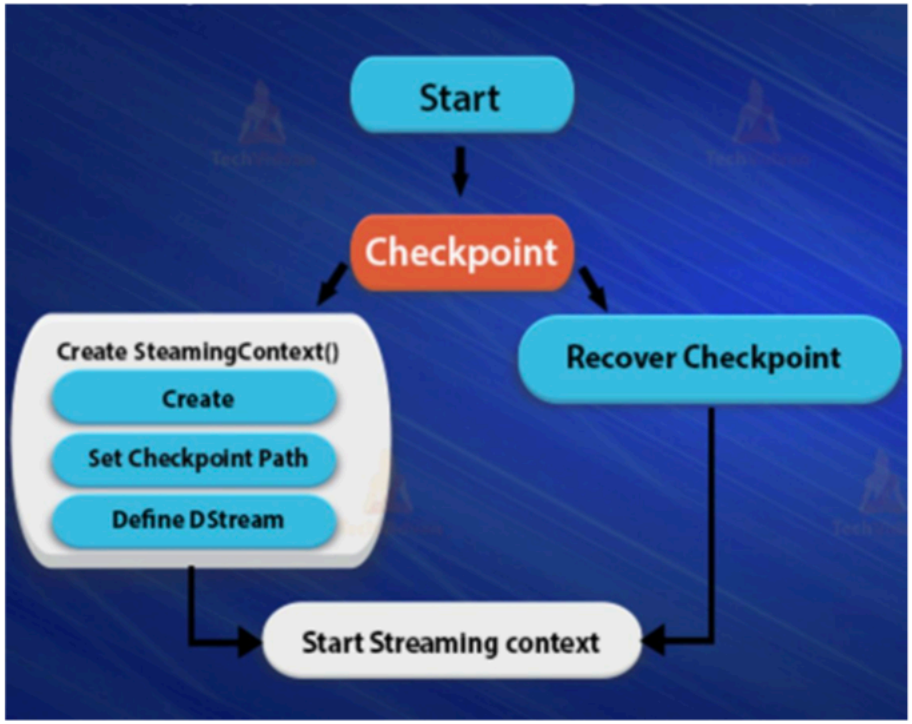


Spark Streaming

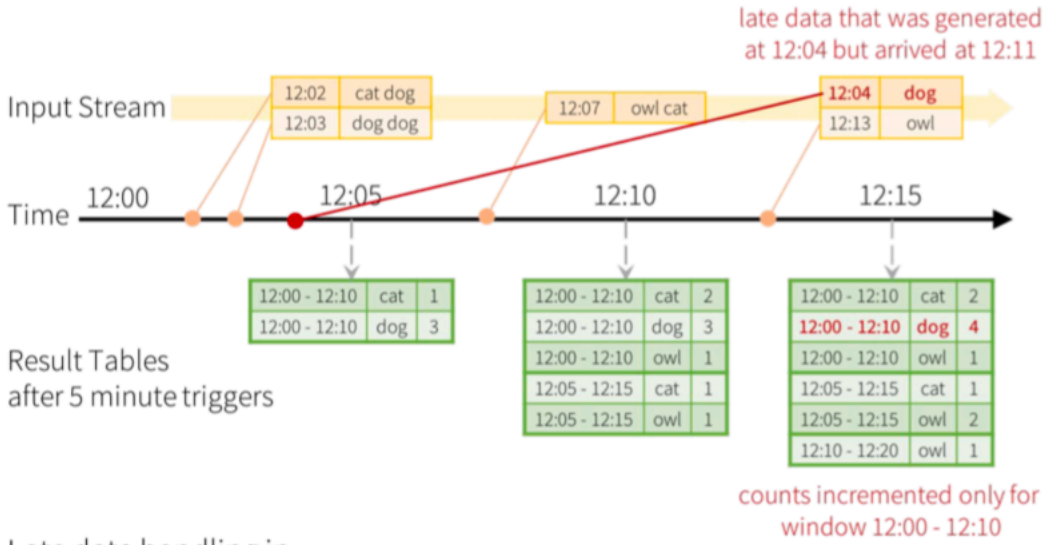


Spark Streaming використовує батч інтервали для обробки подій даних

Відновлення після падіння

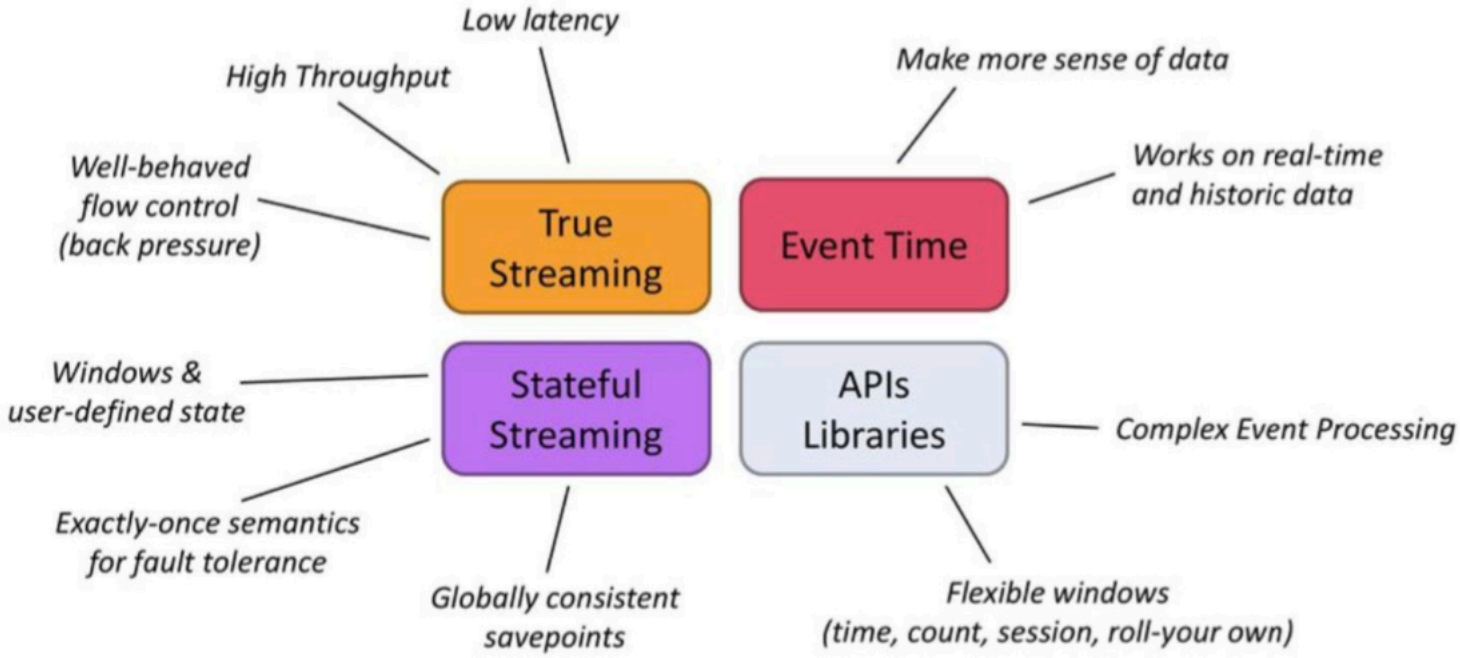


Windowing



Late data handling in
Windowed Grouped Aggregation

Apache Flink



Spark vs Flink

Apache Flink

Spark is not a true real time processing. It is Near to real time processing framework.	Flink is True processing framework.
At heart Spark is a Batch processing framework.	At heart Flink is a Stream processing framework.
Spark's streaming computation model is based on Microbatching.	Flink's Streaming model is based on Windowing and Checkpointing.
Spark is implemented in Scala.	Flink is implemented in Java.
Spark does not have an efficient memory manager.	Flink has it's own efficient automatic memory manager.
Frequently gets Out of Memory.	

Висновки

- Було досліджені алгоритми обробки потоків даних у режимі близькому до реального часу в хмарному середовищі із застосуванням паралельних та розподілених обчислень
- Були дослідження способи забезпечення відмовостійкості при обробці потоків даних та способи обробки даних що приходять із запізненням з використанням Big Data екосистем Spark, Flink та Kafka