Unit 5-
MINING DATA STREAM

When do you want to know ?

or

Later

Now ?

- Whether individual or Business
- Important things are always happening NOW
- Maximize data value process and act in real time

Real-time insight preserves or creates value

OpsClarity Report Summary:

- 92% plan to increase their investment in stream processing applications in the next year
- 79% plan to reduce or eliminate investment in batch processing
- 32% use real time analysis to power core customer-facing applications

http://info.opsclarity.com/2016-fast-data-streaming-applications-report.html

Businesses, crave ever more timely data, and switching to streaming is a good way to achieve lower latency.

*

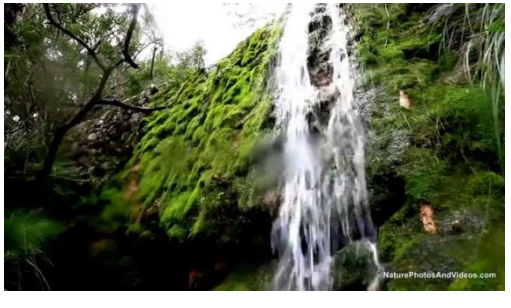
Data Stream

Data stream [7] is nothing but sequence of data objects with respect to time and can be ordered pair (S, T) where:

- S is a sequence of tuples and
- T is a sequence of positive real time intervals.

There are some typical characteristics of data streams:

- Continuous arrival of data objects
- Disordered arrival of data objects
- Potentially unbounded size of a stream



Techniques Stream Data Mining

- **Sampling**
- **Sketching**
- **Load Shedding**
- **Synopsis Data Structures**
- **Aggregation**
- **Sliding Window**

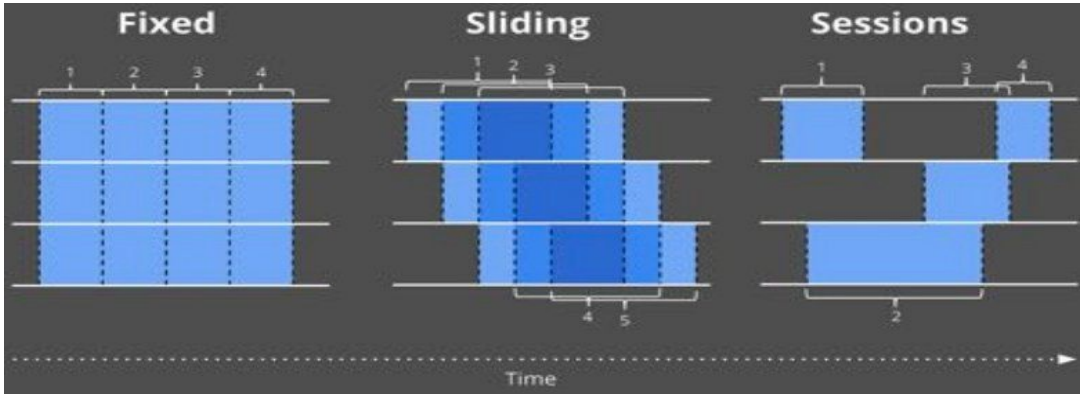


Figure: Windowing Techniques for Stream Data

Stream Bigdata Mining

- "V" (Volume, Velocity, and Variety)

- **Sampling**
- **Clustering**
- **Compression**
- **Wavelets**
- **Histogram**
- **Micro-clustering**

Lambda Architecture

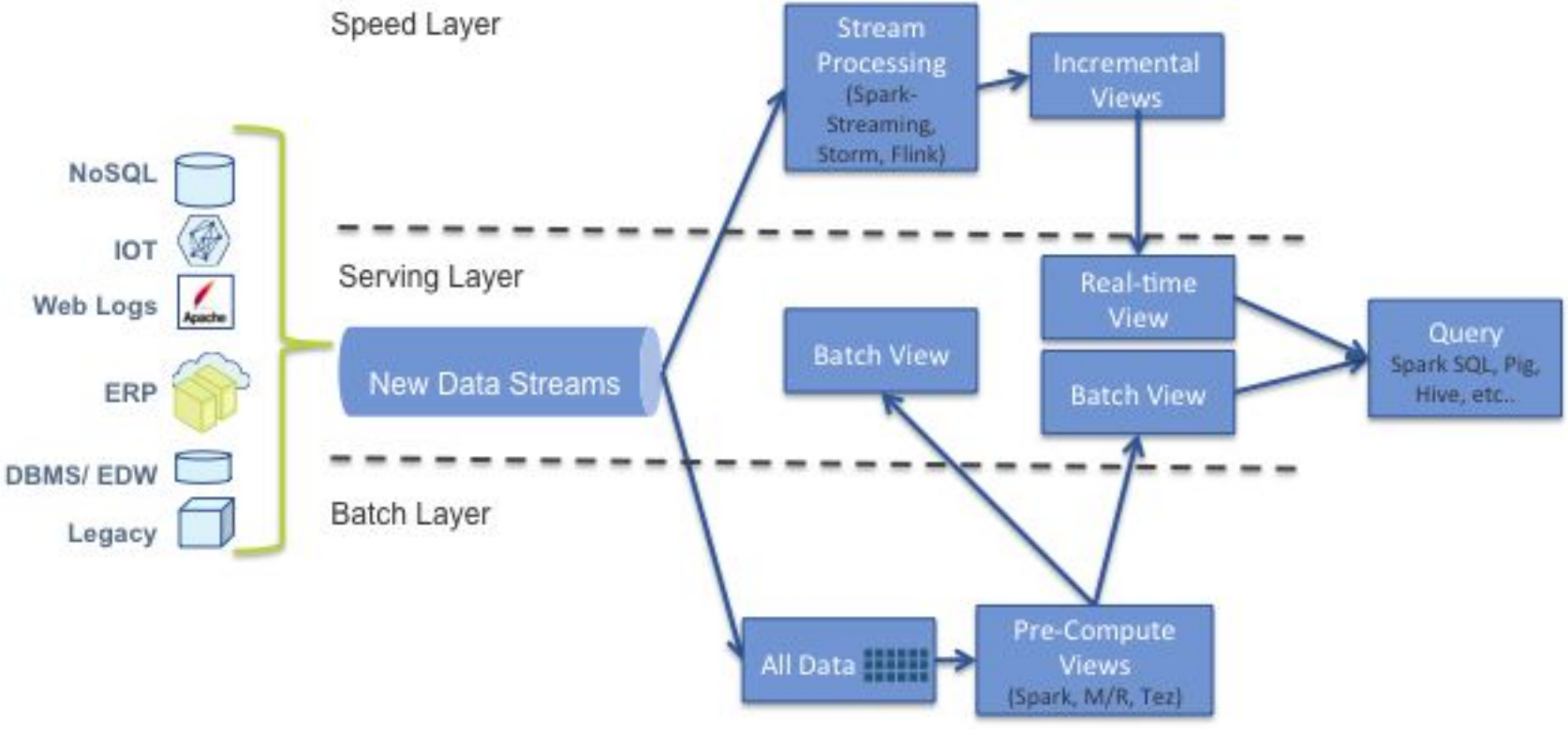


Figure : Lambda Architecture [42]

Lambda Architecture—Requirements

- **Fault-tolerant** against both hardware failures and human errors
- Support variety of use cases that include **low latency** querying as well as updates
- **Scalability**
- Extensible, so that the system is manageable and can accommodate **newer features** easily

Lambda Architecture—Layers

- **Batch layer**

  Managing the master dataset, an immutable, append-only set of raw data

- **Serving layer**

  Indexes batch views so that they can be queried in ad hoc with low latency

- **Speed layer**

  Accommodates all requests that are subject to low latency requirements. Using fast and incremental algorithms deals with recent data only

Applications

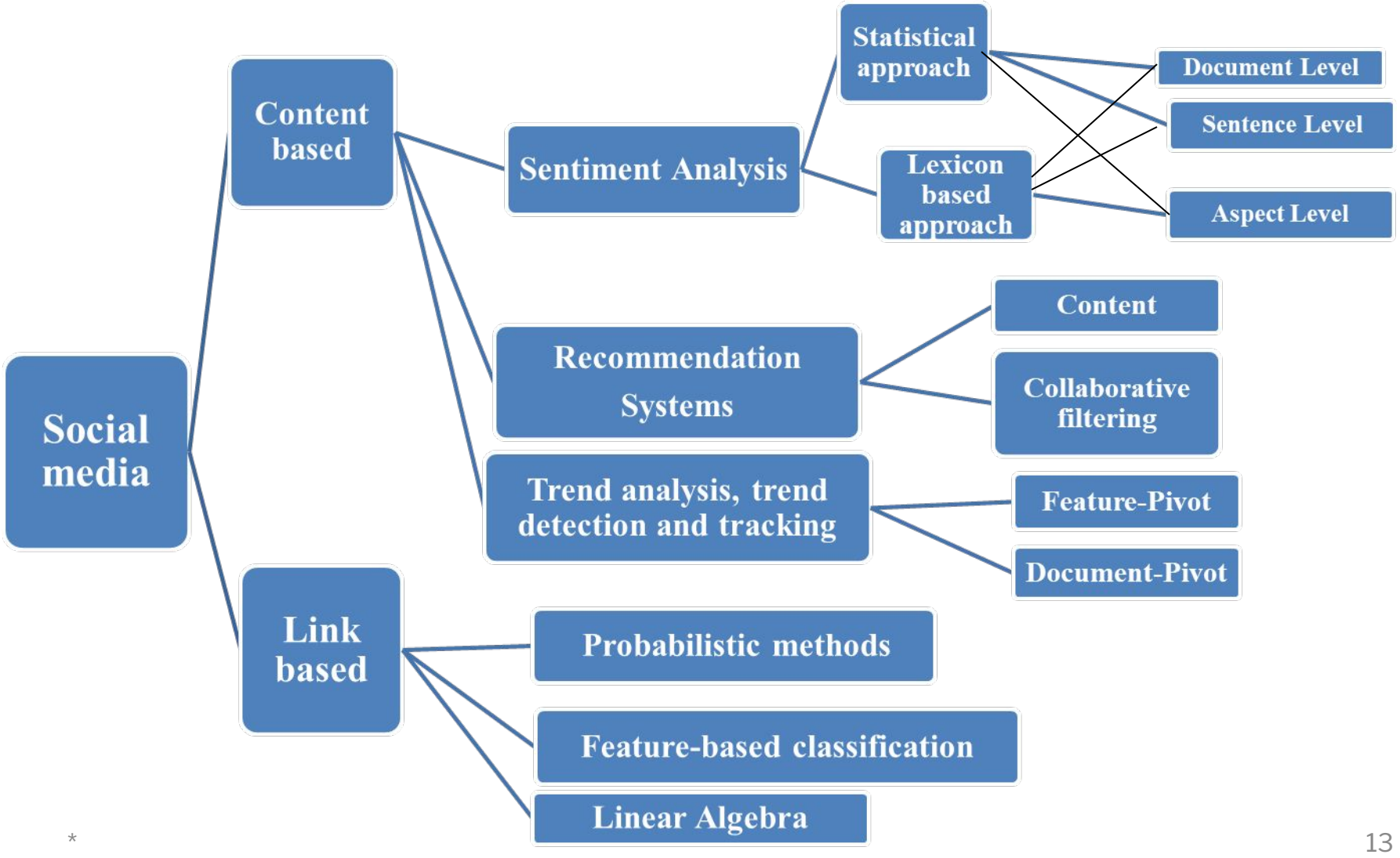
- Intelligent transport Systems
- Stock market
- Network monitoring
- Intelligence and surveillance
- E-commerce
- Social Media

Social Media Data

- **Social media is defined as a group of Internet-based applications that allow the creation and exchanges of user-generated content.**
- Social media gives users an easy-to-use way to communicate and network with each other on an unprecedented scale.

Kaplan, Andreas M., and Michael Haenlein. "Social media: back to the roots and back to the future." *Journal of Systems and Information Technology* 14.2 (2012): 101-104.

Techniques for Social Media Mining



Architecture

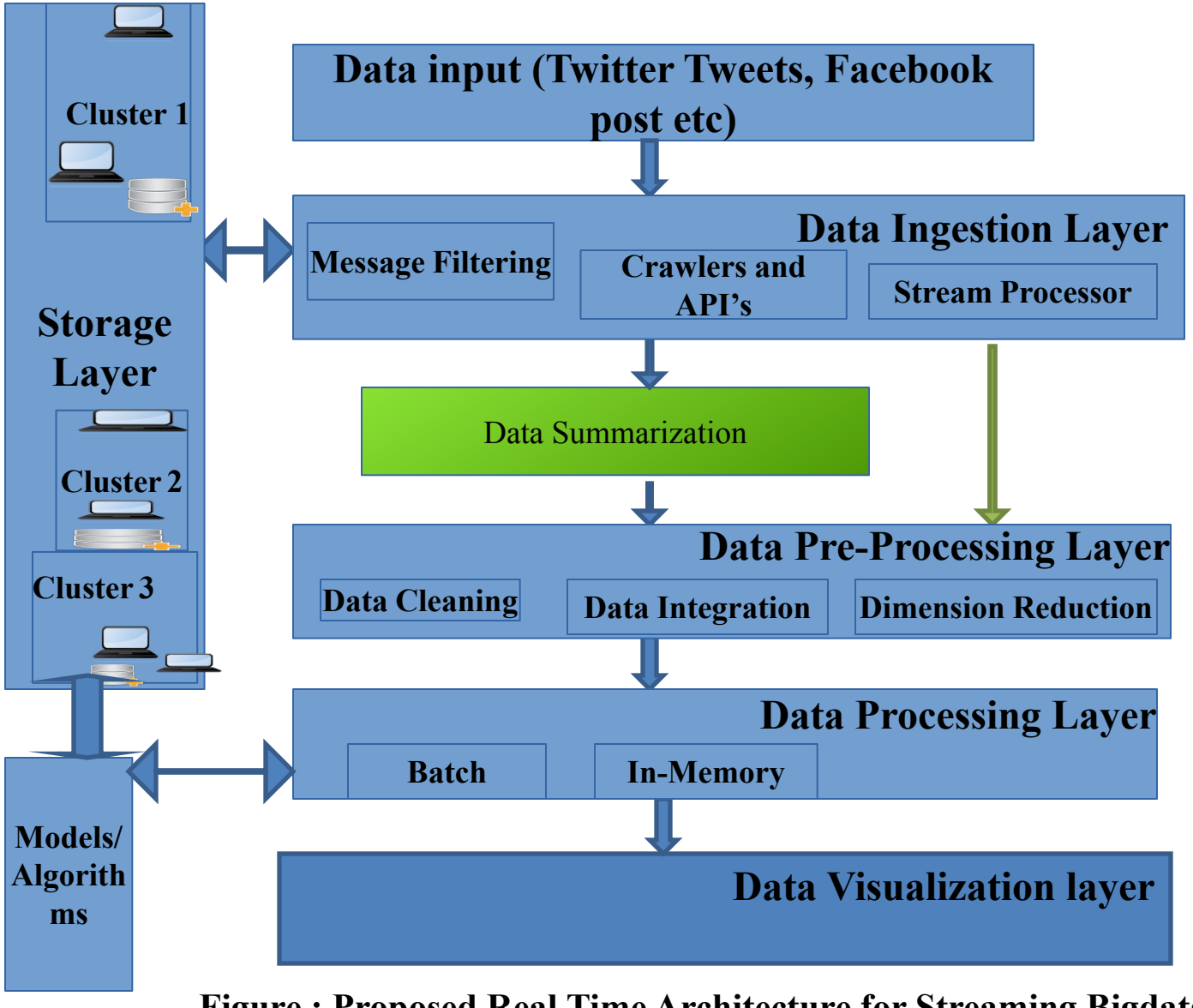


Figure : Proposed Real Time Architecture for Streaming Bigdata

Table : Review of Real Time Platforms for Stream Bigdata

System/ Tools	Processing Model	Stream type	Operating system	Open source	Built in language	Supportive Langauages	Current release/ version	Developed at	Available on	Function	Features
Storm	Batch and Real time	Tuples	OS independant	Yes	Clojure	Any language	0.9.6	BackType	Apache Software Foundation	Distributed real-time computation	Secure Multi-Tenant Deployment
Flink	Batch and Real time	Strings	OS independant	Yes	Java and Scala	Java, Scala, and Python	0.10.2	DataArtisans	Apache Software Foundation	Distributed real-time computation	Low-latency stream processor flexible operator state and streaming windows
Spark	Batch and streaming	discretized stream	Windows and Linux	Yes	Scala	Scala, Java, Python, R	1.6.0	UC, Berkeley	Apache Software Foundation	Large-scale Data Processing	Decentralized hides all cluster management tasks. Checkpointing and recovery minimize state loss
Samza	Batch and Real time	Messages	OS independant	Yes	Scala, Java	CQL, Pig	0.10.0	LinkedIn	Apache Software Foundation	Processing continuous Stream	Simple API Managed state, Fault tolerance Durability, Pluggable Processor isolation
Kafka	Real time	Messages	OS independant	Yes	Scala	Python,Go, C/C++,.net, Clojure, Ruby	0.9.0.1	LinkedIn	Apache Software Foundation	Data integration	Distributed by Design Fast Scalable Durable

*

Summary

- Bigdata needs extraordinary techniques to efficiently process large volume of data within limited run times
- Social media growth and diversity have profoundly affected how people process and interpret new knowledge
- Architecture for real time stream Bigdata has been proposed
- Investigated into the current popular platforms and techniques that can be useful in implementing real-time systems for Bigdata
- Plan to implement and evaluate the proposed approach to perform sentiment analysis with real-time data stream on social networking sites like Twitter

- References as per report

Publication

Bharat Tidke and Rupa Mehta, " *A Comprehensive Review and Open Challenges of Stream Big Data,*" **Springer** International Conference on Soft Computing: Theories and Applications 27-29 December 2016 Jaipur, India.
(Accepted)

Thank You