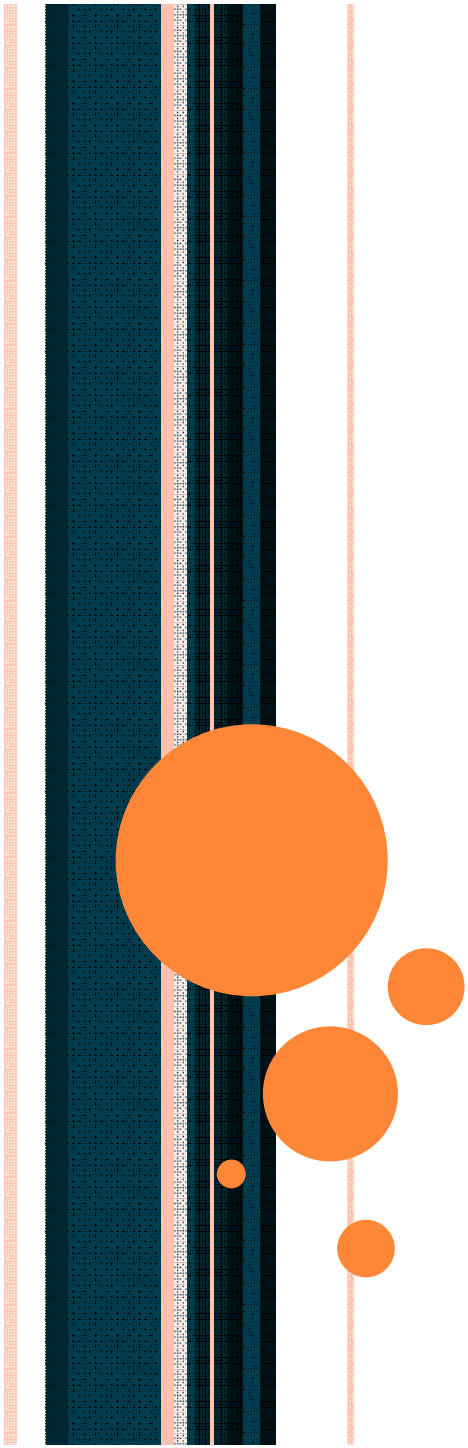


REAL TIME ANALYTICS PLATFORM

REAL TIME ANALYTICS

- the use of, or the capacity to use, all available enterprise data and resources when they are needed.

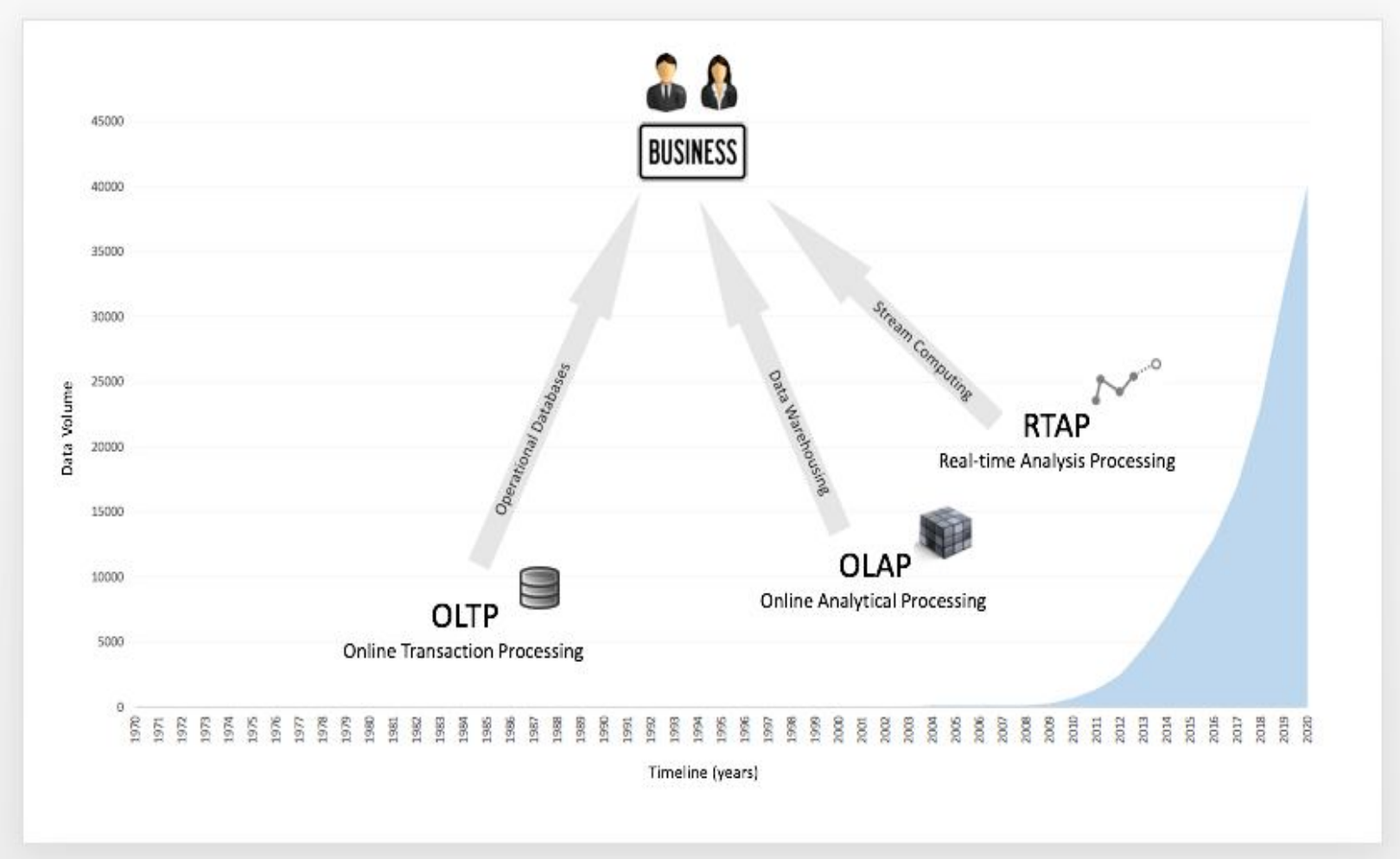
- consists of dynamic analysis and reporting, based on data entered into a system less than one minute before the actual time of use.

- also known as
  - real-time data analytics,
  - real-time data integration,
  - real-time intelligence.

The evolution of analysis: OLTP, OPAL, RTAP

- Big Data and IoT brings new challenges but also new opportunities everyday.
- There are multiple ways to store, process and analyze data.
- Reports were earlier built on Online Transaction Processing(OLTP) databases.
- Then we moved to a more intelligent analysis using Online Analytical Processing(OLAP).
- But now, given the huge volume, variety and velocity of data, we need to analyze it faster.
- This is why Real-Time Analysis Processing (RTAP) came into picture.

RTAP EVOLUTION CHART



SUPPORTING TECHNOLOGIES

- Processing In Memory(PIM)
- In-database Analytics
- Data Warehouse appliances
- In-memory Analytics
- Massively Parallel Programming(MPP)

APPLICATIONS

- Customer relations management,
  - real-time analytics can provide up-to-the-minute information about an enterprise's customers and present it so that better and quicker business decisions can be made -- perhaps even within the time span of a customer interaction.
  - can support instant refreshes to corporate dashboards to reflect business changes throughout the day.

- In data warehouse:
  - real-time analytics supports unpredictable, ad hoc queries against large data sets.

APPLICATIONS

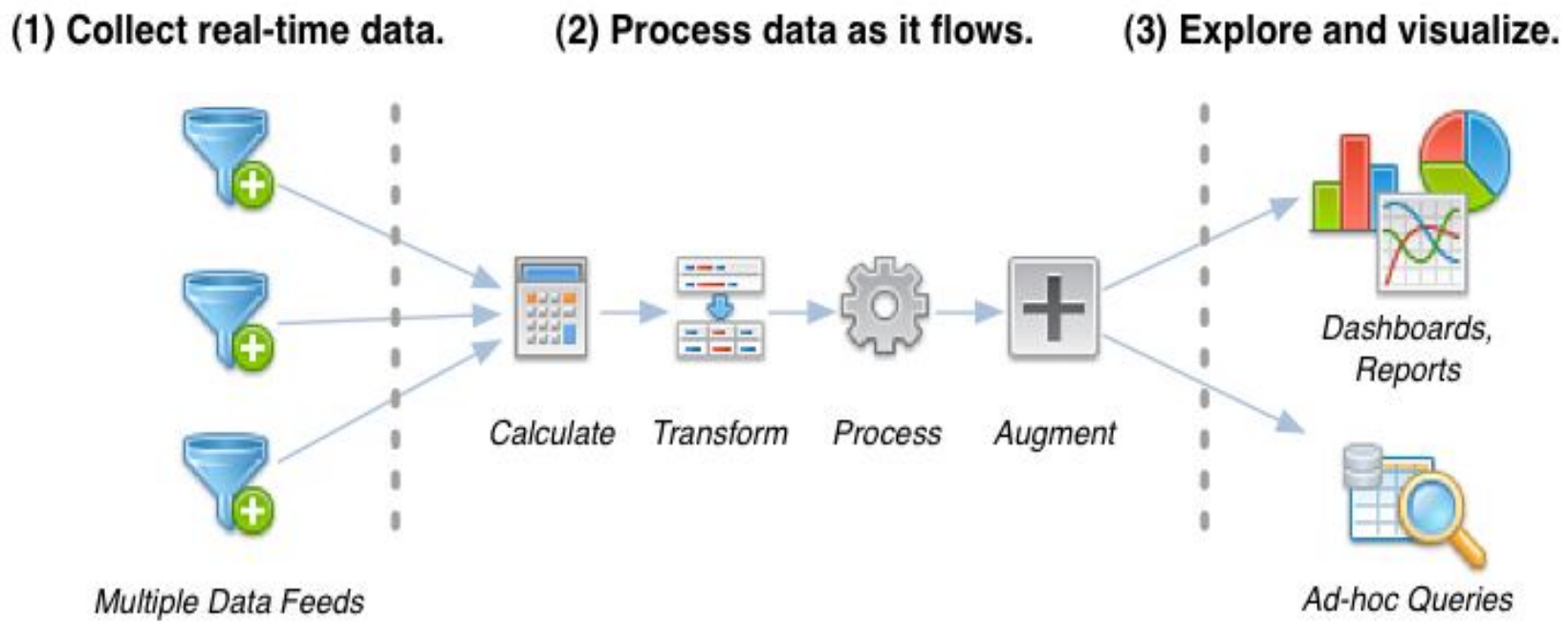
- Scientific analysis
  - such as the tracking of a hurricane's path, intensity, and wind field, with the intent of predicting these parameters hours or days in advance.

- The adjective real-time refers to a level of computer responsiveness that a user senses as immediate or nearly immediate, or that enables a computer to keep up with some external process (for example, to present visualizations of Web site activity as it constantly changes).

RTAP

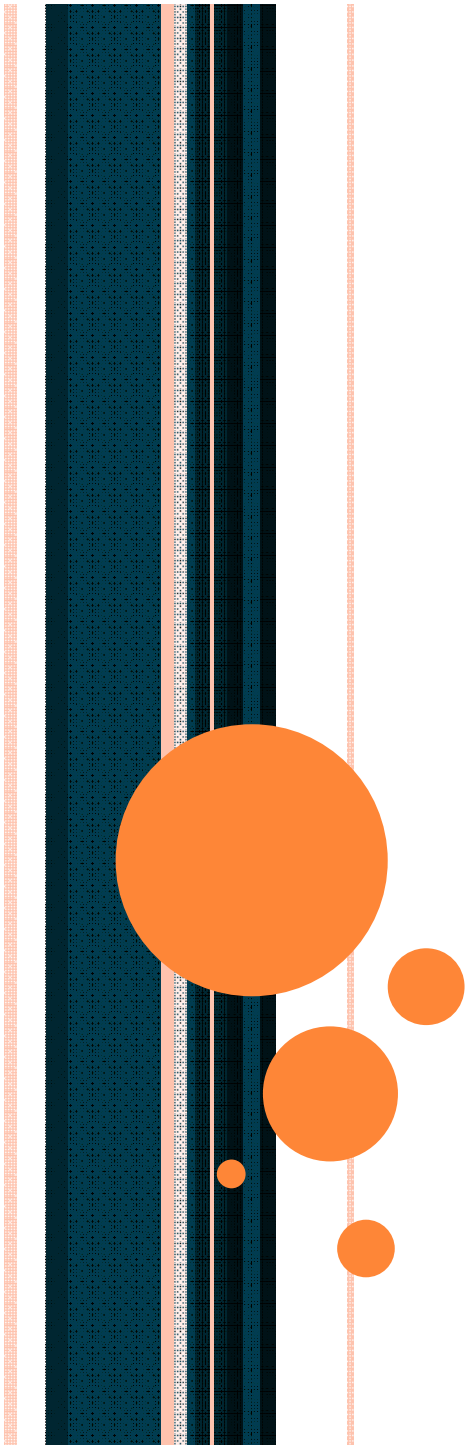
- Acronym for Real Time Analytics Platform
- Enables organizations to make the most out of real-time data by helping them to extract the valuable information and trends from it.
- Such platforms help in measuring data from the business point of view in real time, further making the best use of data.

GENERIC ARCHITECTURE FOR A RTAP



EXAMPLES OF RTAP

- **Piwik** : Piwik offers real time web analytics reports.
- **MemSQL**: in-memory database and real-time analytics platform, which enables organizations to perform real-time analytics at Big Data scale. Companies can manage the velocity of Big Data transactions without sacrificing their ability to interact with and analyze that data via SQL
- **MongoDB**: The City of Chicago cuts crime and improves citizen welfare with a real-time geospatial analytics platform called WindyGrid. Using MongoDB, it analyzes data from 30+ different departments – like bus locations, 911 calls, and even tweets – to better understand and respond to emergencies.
- **StreamAnalytix** : designed to rapidly build and deploy streaming analytics applications for any industry vertical, any data format, and any use case.



Opinion mining (sentiment analysis)

INTRODUCTION

- Aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document.
- Uses Natural Language Processing (NLP),text processing, text analysis and computational linguistics to identify and extract subjective information from the source material.
- Widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service.

DATA PREPROCESSING

- A vital step in the process
- Extract the relevant content from the input data
- It improves the classification accuracy and decrease the running time of algorithm.

- Message String → feature vector

- Examples:
  - Replacing emotions: replace with either FROWN or SMILE keyword
  - URL extraction
  - Identification and removal of punctuations
  - Removal of Stop Words
  - Stemming

Approaches for sentiment classification

- Supervised approach

  Example: SVM, Naïve Bayes.

- Unsupervised approach

  Example: NLP Patterns.

- Semi Supervised approach

  Example: Lexicon+Classifier.

FEATURES

- Semantic Features
  - Leverage meaning of words
  - Can be done manually/semi/fully automatically.
  - Example: If a phrase has better association with the word “excellent” than with “poor” it is positively oriented and vice versa.
- Syntactic Features
  - Usage of principles and rules for constructing sentences in natural language.
  - Example: a sentence containing an adjective and “!” could indicate existence of an opinion.

FEATURES

- Linked-based Features
  - Use link/citation to determine sentiments of documents.
  - Opinion Web pages heavily linking each other often share similar sentiments.
  - Not a popular approach.
- Stylist Features
  - Incorporate stylometric/authorship studies into sentiment classification.
  - Style markers have been shown highly prevalent in Web discourse.

EVALUATION

- The accuracy of a sentiment analysis system is, in principle, how well it agrees with human judgments.
- This is usually measured by:
  - Precision

    measures the exactness of a classifier. A higher precision means less false positives, while a lower precision means more false positives.
  - Recall

    Recall measures the completeness, or sensitivity, of a classifier. Higher recall means less false negatives, while lower recall means more false negatives. Improving recall can often decrease precision because it gets increasingly harder to be precise as the sample space increases.

EVALUATION

- F1 SCORE:

- The two measures – Precision and Recall are combined to give the F1 score as:

$$F1 = 2 * Precision * Recall / (Precision+Recall)$$

CHALLENGES

- Classifying the polarity of a given text .
- Types
  - Positive
  - Negative
  - Neutral
- Positive opinion

  If a customer said a laptop's battery life was long.
- Negative opinion

  If a customer said that the laptop's start-up time was long.
- These differences mean that an opinion system trained to gather opinions on one type of product may not perform well for another

CHALLENGES

- Most of the traditional text processing relies on the fact that small differences between two pieces of text don't change the meaning very much.

- In opinion mining, however

  "The movie was great"

  is very much different from

  "The movie was not great"

CHALLENGES

- Contradictory statements.
- Reviews may both positive and negative comments because people are more likely to combine different opinions in the same sentence.
- “The movie bombed even though the lead actor rocked it”.
  - This is easy for a human to understand, but more difficult for a computer to parse.
- “That movie was as good as his last one.”
  - This is entirely dependent on what the person expressing the opinion thought of the previous film.

CHALLENGES

- Subjectivity / Objectivity identification.
  - More difficult than polarity identification.
  - The subjectivity of words and phrases may depend on their context.
  - Objective document may contain subjective sentences.

    Example: “ News article quoting people’s opinion.”
  - Results are largely dependent on the definition of subjectivity used when annotating texts.
  - Experiments have proved that removing the objective sentences from a document before classifying its polarity helped improve performance.

CHALLENGES

- Detection of spam and fake reviews
  - A website may contain both authentic and spam contents.
  - For effective sentiment analysis, this spam content should be eliminated before processing.
  - Technique used: Identifying duplicates by Outlier Analysis and by considering reputation of reviewer.

APPLICATIONS

- Purchasing Product or Service.
- Quality improvement in Product or Service.
- Marketing Research.
- Recommendation System.
- Opinion spam detection.
- Decision making.
- Policy marketing.
- Detection of “flame”
  - Automatic detection of arrogant words, over heated words or hatred languages used in emails or forums entries or tweets on various internet sources.

THANK YOU!!