Scalable Online Analytics for IoT Applications using Big Data Platforms

Arun Verma
Indian Institute of Technology Bombay
Mumbai, India
v.arun@iitb.ac.in

Yogesh Simmhan
Indian Institute of Science
Bengaluru, India
simmhan@cds.iisc.ac.in

Nandyala Hemachandra
Indian Institute of Technology Bombay
Mumbai, India
nh@iitb.ac.in

Introduction

- Internet of Things (IoT) is rapidly penetrating every sphere of life. By 2020, 6.1 billion smartphones and over 50 billion connected devices [1] in the world will generate data.
- In the Indian context, emerging Smart Cities offer infrastructure services such as smart water, power and transportation management based on these IoT technologies (Figure 1).

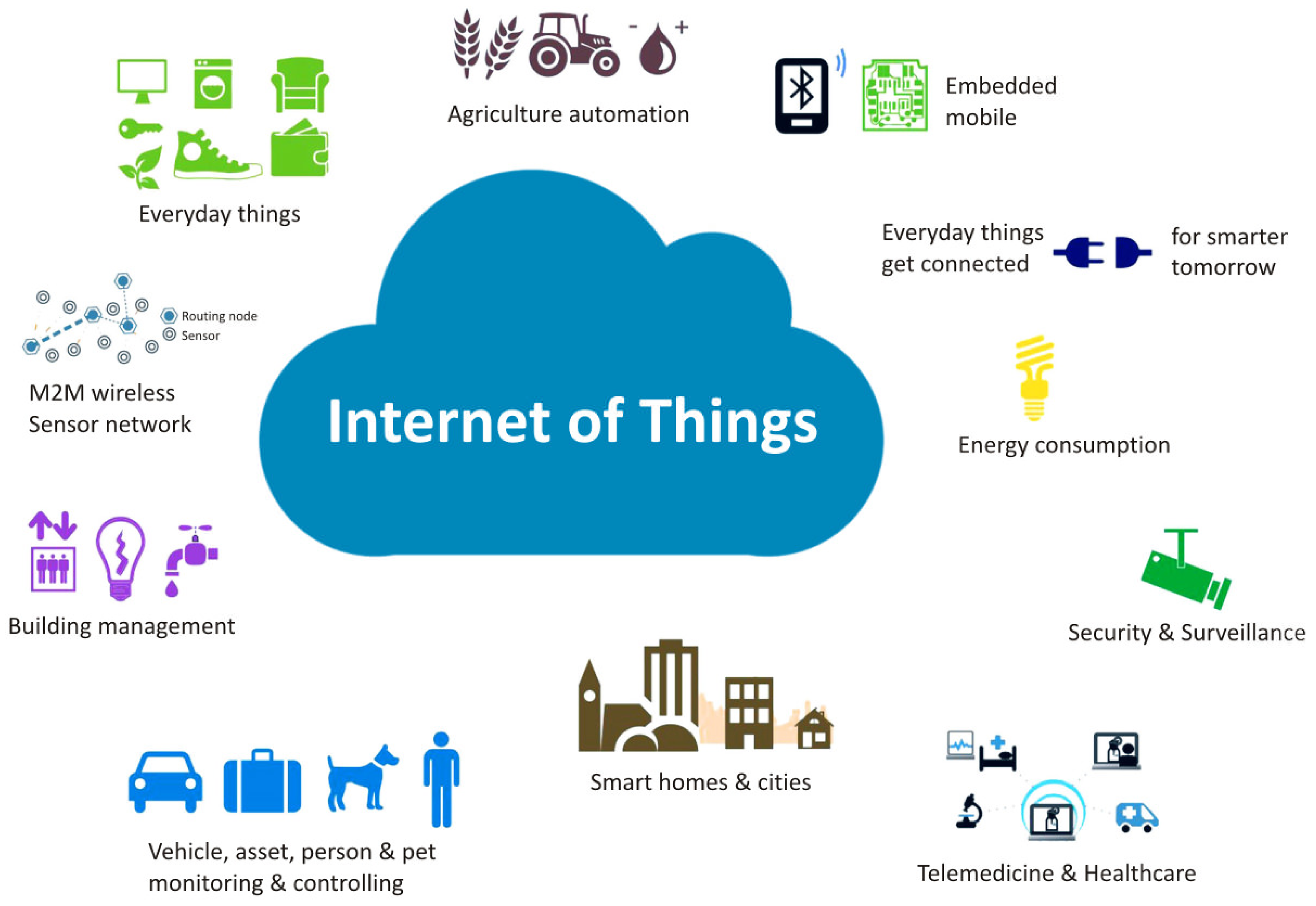

Figure 1: Internet of Things (IoT) Applications [2]

- Big Data platforms allow for scalable and distributed processing of such large and fast datasets, but need to be integrated with and extended to support the unique needs of IoT applications.
- Distributed software platforms for managing such Big Data exist in the form of Hadoop and Spark for large volumes, and Storm and Flink for fast data streams.

Key gaps in leveraging these platforms

- Integrating the capabilities of Big Data platforms for a scalable end-to-end distributed IoT software framework.
- Developing techniques for online analytics and real-time decision-making needs of these Big Data applications.

Goal : To develop an end-to-end distributed framework to support the IoT applications.

Streaming Data Applications Overview

The end-to-end streaming data applications (Figure 2) can be categorized into three parts: **Input, Processing and Output.**

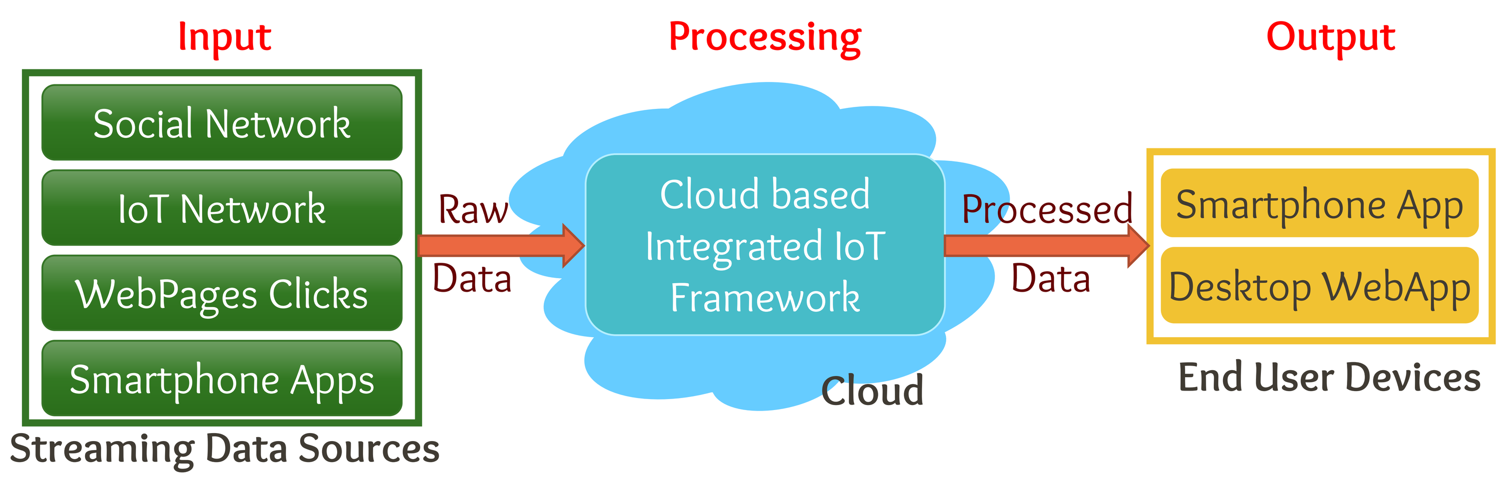

Figure 2: End-to-end IoT Applications.

Integrated IoT Framework Implementation

We have developed an integrated IoT framework for the Smart Campus project at IISc [3] for sustainable water management. This framework is designed in an extensible manner to support other domains such as smart power and renewable energy [4]. The architecture of the IoT framework (Figure 3) is divided into two parts: ***Stream Processing Pipeline*** to support real-time processing of incoming data streams, and ***Analytics Pipeline*** for analyzing the archived data.

Stream Processing Pipeline

- Raw data streams are formatted and published by applications of the first type that run on sensors or edge devices like smart phones and Raspberry Pi.
- Publish-subscribe message broker (Apache Apollo) on the Cloud that exposes topics where the observations are published using the MQTT protocol.
- A distributed stream processing system (DSPS; using Apache Storm) receives the data streams by subscribing to the specific topics in message broker and processes the incoming messages.
- Streaming engine executes data preprocessing and information integration pipelines that cleans the data, does formatting and unit conversions, interpolation and smoothing, and annotations. These quality-checked and enriched data streams along with the original raw streams are then stored into a NoSQL database (Apache HBase).
- Processed streams are also published on topics with the broker for downstream users and applications to consume in real-time.
- Depending on the application, the data streams can arrive at 1000's of messages/sec, and the broker, DSPS and NoSQL database are designed and integrated to scale to this.

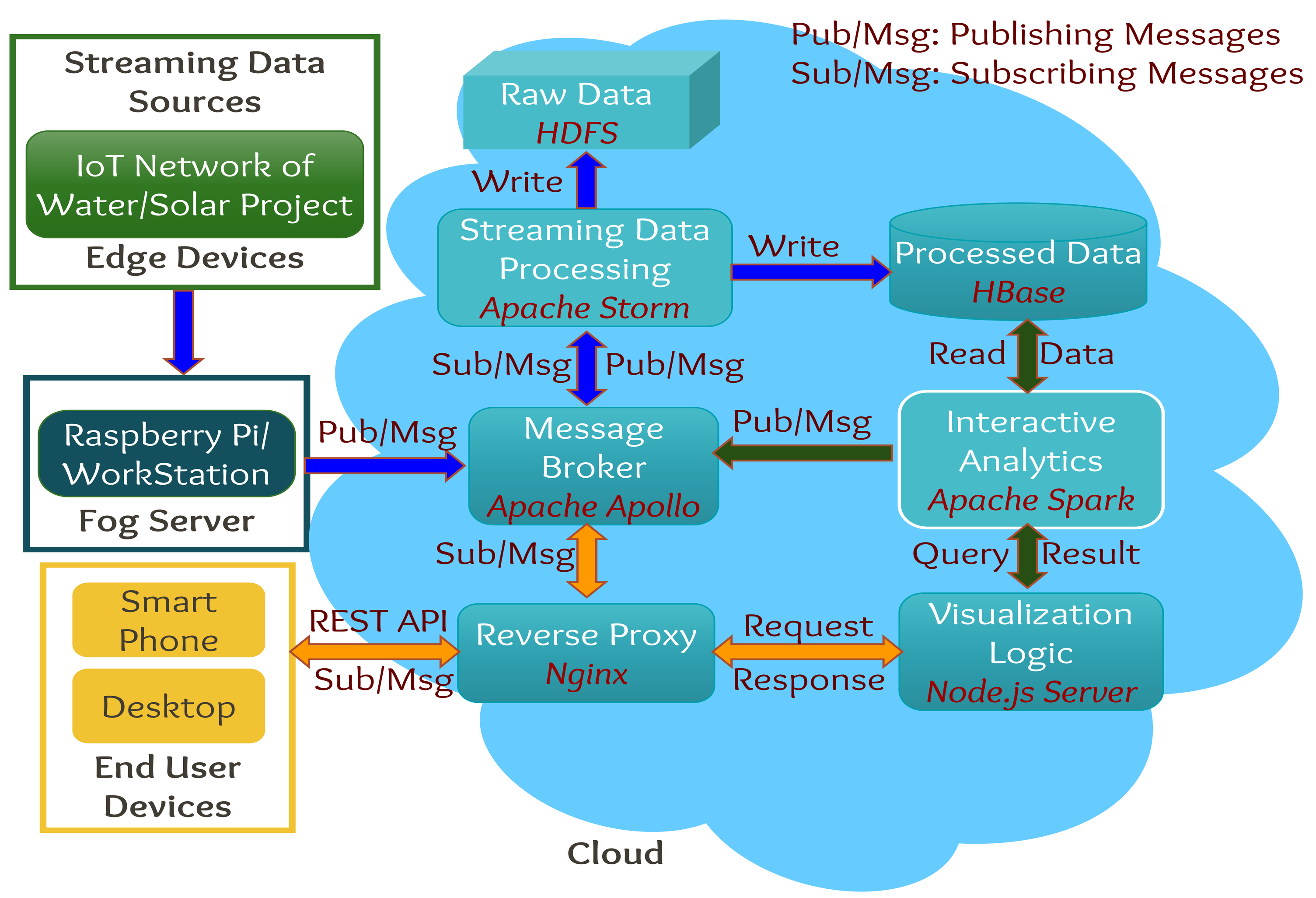

Figure 3: Architecture of the end-to-end framework for the Smart Campus project. Blue lines represent the *stream processing pipeline*, green lines the *offline analytics pipeline* and orange lines are shared by both.

Big Data Analytics Pipeline

- Processed data has to be analyzed to offer predictions, identify patterns and drive decisions in the IoT infrastructure, such as water pumping operations or for demand-response from power consumers [5].
- Current framework supports basic offline analytics and online visualization using D3.js which plugs into the broker topics.
- Offline analytics operates on the data archive of Gigabytes-Terabytes in size and uses Apache Spark which offers a fast, scalable and distributed engine for processing such large-scale data.
- We use it for simple statistical analytics such as max, min, average, etc., for a given time window that is launched and executed in a batch mode from a portal. The responses are returned in a synchronous or asynchronous mode using the broker.

Future Work: Batch Updated Online Learning

Motivation

- There is a need for predictive models that can operate over data streams as they arrive for low latency forecasts and classification that can be used for better decision making. Further, the prediction models themselves need to be retrained to reflect current conditions.
- For example, in the Smart Campus project, the prediction of the water usage by campus residents can be used to find a suitable time to pump water to overhead tanks. Besides modeling, using historical data also captures some of the daily conditions defined by events on campus and schedule of the city utilities. Further, the dynamic pricing of electricity implies that the cost for pumping, which is non-trivial, has to be accounted for to ensure we capitalize on low pricing periods.

Online Learning Model

- Batch-updated online learning module (Figure 4) is divided into two parts: Online predictions and updating the online model.
  - **Online Prediction :** Predict based on incoming messages using a previously trained online model.
  - **Updating Online Model :** This has a *Mini-Batch Store* to keep the processed data stream for a preset time window, a *Processed Data Store* with historical data, and the *Model Trainer*. The trainer loads the Mini-Batch and Data Store at the end of a time window and after assigning them proper weights, retrains the Online Model based on previous predictions and updates the existing Online Model.
- **Key research challenges :** To develop and integrate novel online machine learning [6] and event mining into the above model, along with suitable evaluation schemes like [7] for classification, that are sensitive to the needs of diverse IoT domains.

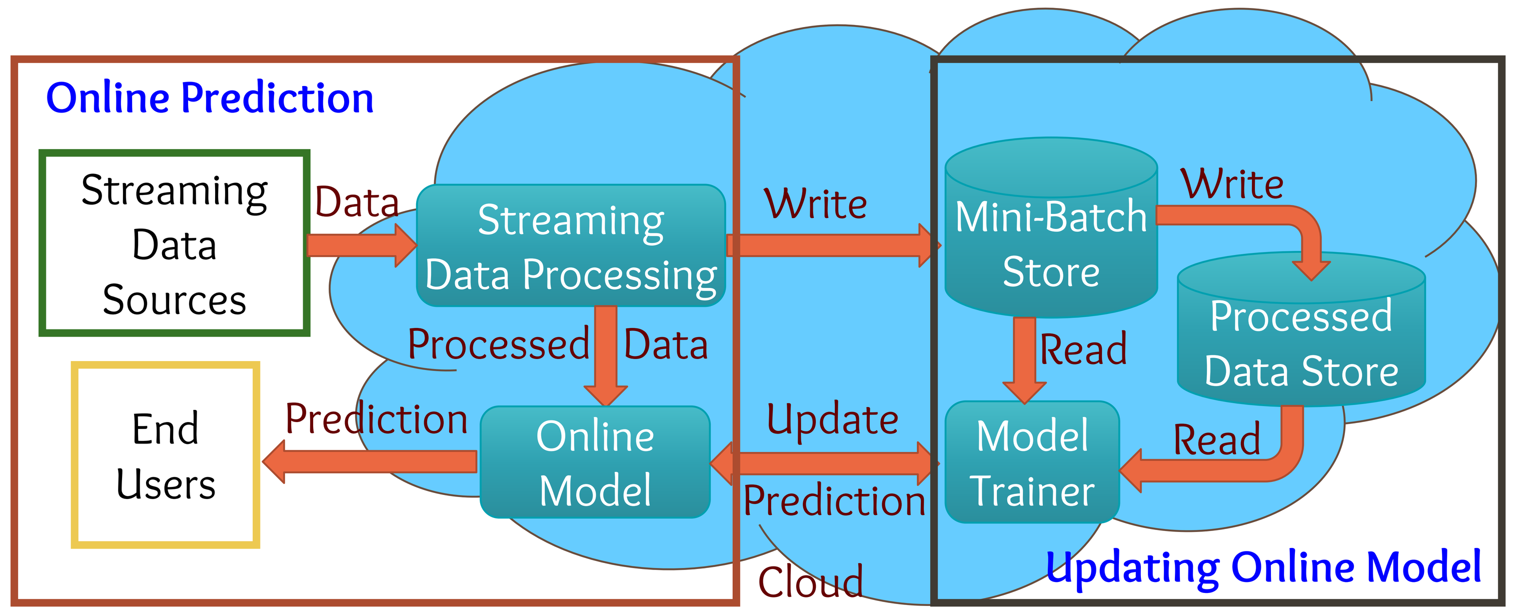

Figure 4: Batch updated Online Learning Module.

Acknowledgments

Smart Campus project is supported by the Department of Electronics & Information Technology (DeitY), Govt. of India and RBCCPS, IISc.

References

[1] Forbes. Big Data: 20 Mind-Boggling Facts Everyone Must Read, Sep 30, 2015. Accessed: 2016-12-28.

[2] Scientech: http://www.scientechworld.com/internet-of-things/iot-solutions/iot-builder. Accessed: 2017-03-10.

[3] Smart Campus Project: http://smartx.cloudapp.net/.

[4] Yogesh Simmhan, Saima Aman, Alok Kumbhare, Rongyang Liu, Sam Stevens, Qunzhi Zhou, and Viktor Prasanna. Cloud-Based Software Platform for Big Data Analytics in Smart Grids. *Computing in Science & Engineering*, 15(4):38–47, 2013.

[5] Saima Aman, Yogesh Simmhan, and Viktor K Prasanna. Holistic Measures for Evaluating Prediction Models in Smart Grids. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):475–488, 2015.

[6] Sara Landset, Taghi M Khoshgoftaar, Aaron N Richter, and Tawfiq Hasanin. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, 2(1):1, 2015.

[7] Albert Bifet, Gianmarco de Francisci Morales, Jesse Read, Geoff Holmes, and Bernhard Pfahringer. Efficient Online Evaluation of Big Data Stream Classifiers. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 59–68, 2015.