Towards an Integrated Solution for IoT Data Management

Anderson Chaves da Silva
PhD Candidate at the National Laboratory for Scientific Computing

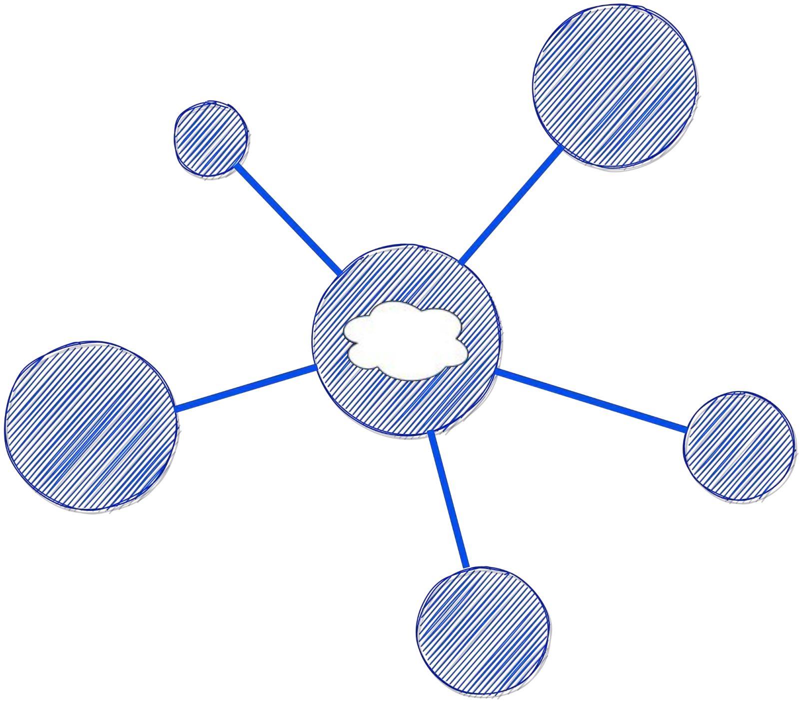
August 16, 2021

The Internet of Things

- IoT Applications: Object Tracking, Anomaly Detection, Real time analysis and decision making

- Domain Specific Services: Smart Home, Smart City, Smart Healthcare, Industry, Sports

- Challenges: Processing and analyzing continuous data streams from heterogeneous networks



IoT Challenges

What are the Challenges?

- ➔ Challenge #1: Large Scale IoT Data Management
  - ◆ Big Range of Heterogeneous Endpoints
  - ◆ Massive Amount of Unstructured Data
  - ◆ Space-Time Correlation

- ➔ Traditional solutions cannot fulfill the requirements of IoT data streams
  - ◆ Quick Response
  - ◆ Scalability
  - ◆ Privacy/Security
  - ◆ Resource Constraints (Memory, Bandwidth, Energy)

IoT Challenges

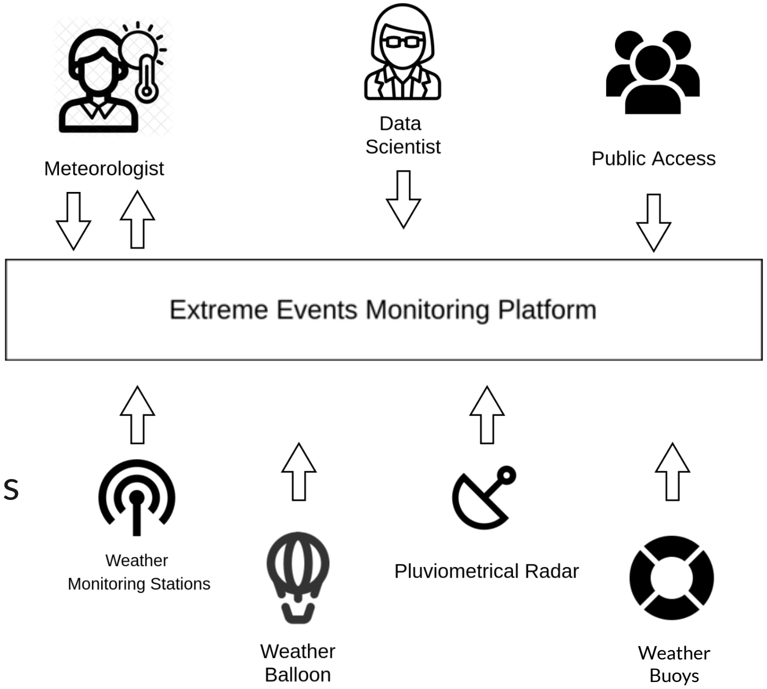
- ➔ Challenge #2: Real-time Analysis
  - ◆ Online Processing x Offline Processing
  - ◆ Lack of integration between the data processing system and ML
  - ◆ Concept Drift

- ➔ Problems...
  - ◆ Need for Light-weight data mining algorithms
  - ◆ Hard to optimize (Query Planning, Lazy evaluation, etc)
  - ◆ ML Models Performance may degrades over time

IoT Challenges

➔ Use case Example
Extreme Weather Events Monitoring Platform

- Multiple Sensor Data Sources
- Streaming data
- Real Time Weather Monitoring and Analysis



IoT Data Management

Actor Oriented Databases

Dynamic Scalability
Asynchronous primitives
Encapsulation

Privacy/Security

Big range of Endpoints

Array Databases:

Array Algebra Operations
Flexible Storage Format

Space-time
correlation

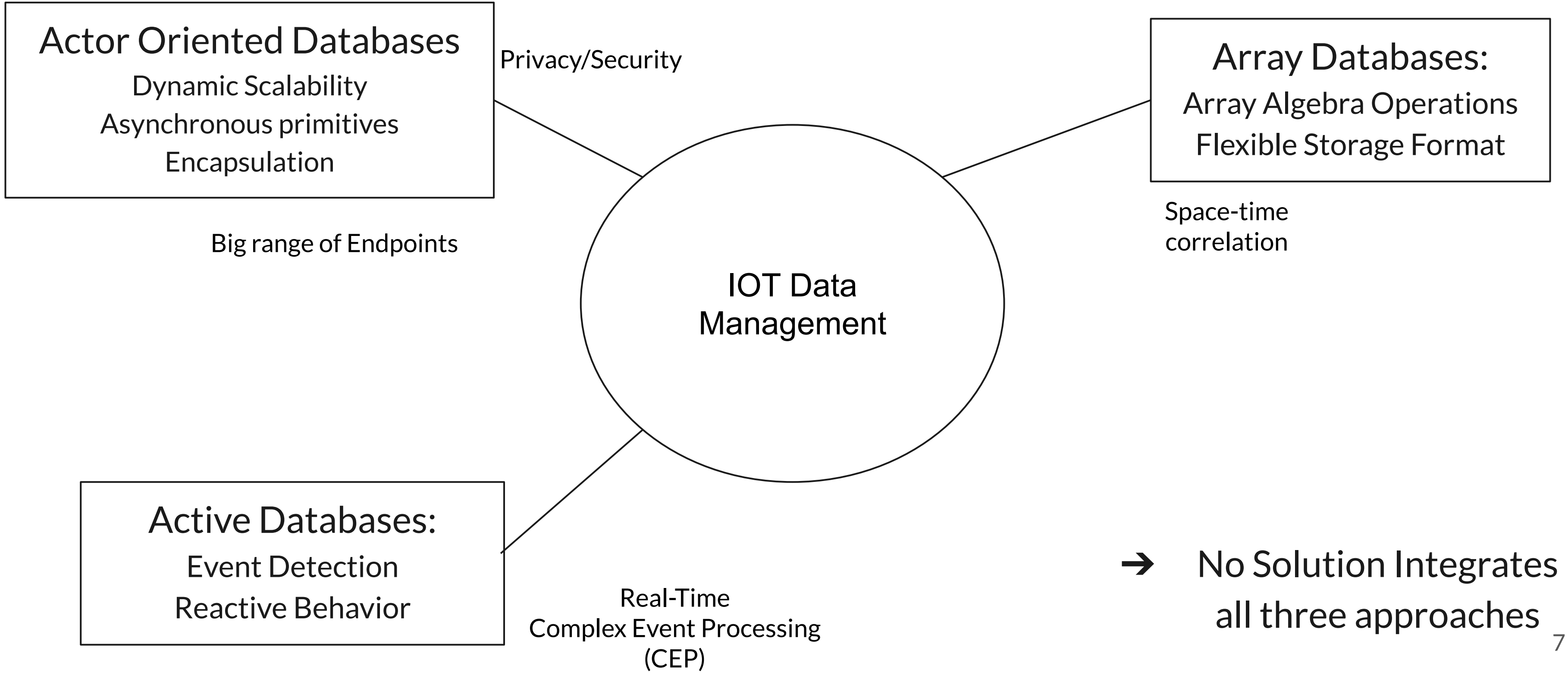
IOT Data
Management

Active Databases:

Event Detection
Reactive Behavior

Real-Time
Complex Event Processing
(CEP)

IoT Data Management

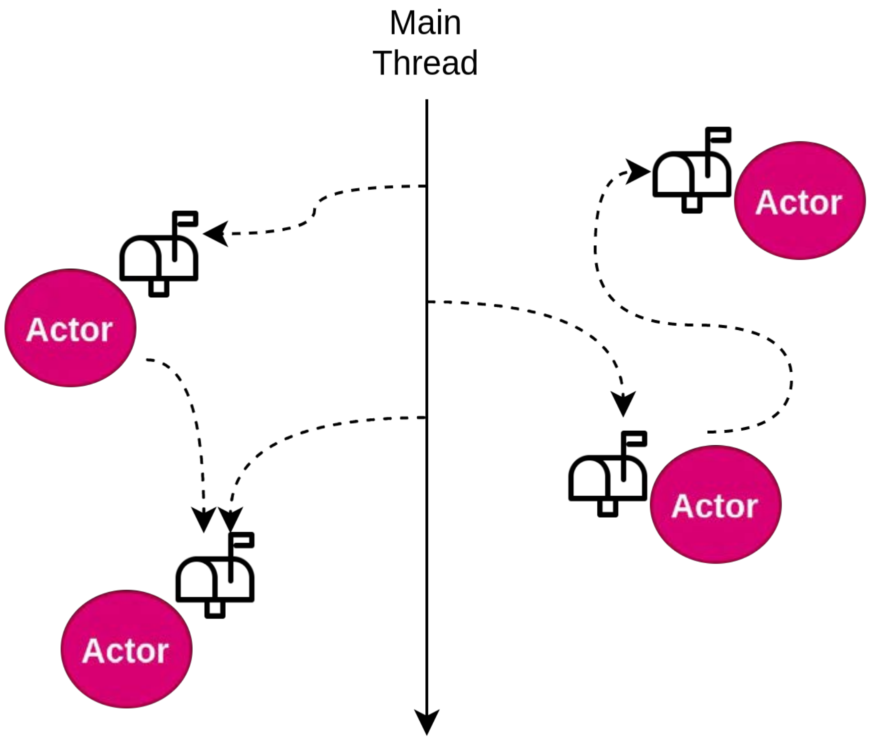


➔ No Solution Integrates all three approaches

IoT Data Management and Actor Oriented Databases

- Actor Oriented Programming
  - Simplifies distributed programming
  - Actors as Fundamental programming unity

- Actors...
  - Messages are queued in the recipient's mailbox
  - No shared-memory state between actors
  - Process one message at a time
  - No multi-threaded execution inside an actor



IoT Data Management and Actor Oriented Databases

- Actor Frameworks, Languages, Toolkits...

ERLANG

RAY

akka

elixir

- Actor Oriented Databases
  - Database -> Actors
  - Actors -> Databases

Orleans

ReactDB

IoT Data Management and Array Databases

- Array Databases provide data analysis based on array algebra
- Especially adequate for multidimensional data applications

  - Implicit Cell ordination: Quick data access

  - No need for indexes on dimensions:
    Smaller storage space

  - Enable data analysis based on array algebra
    operations (e.g. Array Slice)

rasdaman
raster data manager

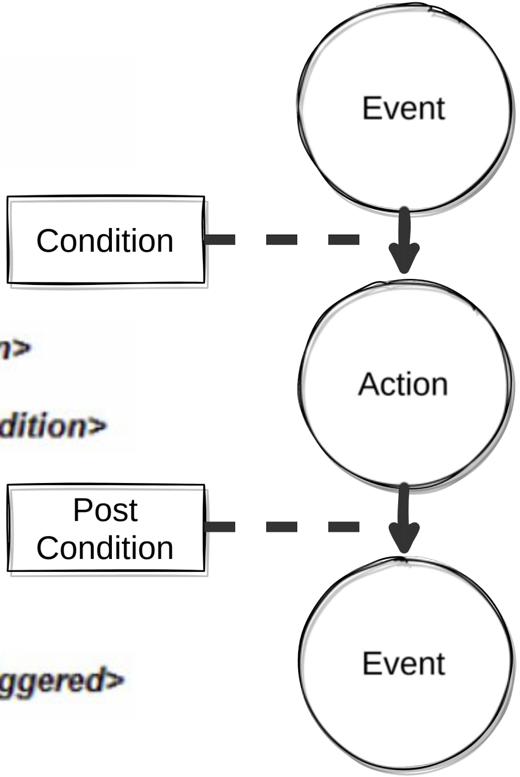
SciDB

IoT Data Management and Active Databases

ON *<Event>*

IF *<Condition>*
DO *<Action>*
Check *<Post condition>*

Trigger *<Event triggered>*



on EVENT if CONDITION do ACTION

- The event of an ECA rule determines when the rule should be evaluated
- The condition of an ECA rule determines whether the action should be executed
- The action of an ECA rule determines how to react if the condition is evaluated true

Active Databases and Complex Event Processing

- Complex Event Processing Techniques
  - Extend the Logic behind ECA Rules
  - Real-time stream processing for monitoring and detection of arbitrarily complex patterns in massive data streams
  - Each data item is abstracted as an event produced by a data source
  - Multiple simpler events combined to produce more complex ones, that match previously defined patterns

Concept Drift Problem

Formal Description:

➔ Concept drift: a phenomenon in which the statistical properties of a target variable change over time in an arbitrary way

Concept Drift Problem

Formal Description:

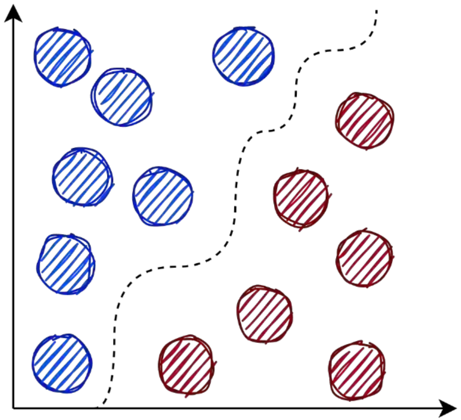
- ➔ Concept drift: a phenomenon in which the statistical properties of a target variable change over time in an arbitrary way
- ➔ Reasons: Changes in hidden not measured variables
- ➔ Must be taken into consideration in an efficient IoT Data Platform

Concept Drift Problem

Formal Description:

➔ Concept drift: a phenomenon in which the statistical properties of a target variable change over time in an arbitrary way

➔ Reasons: Changes in hidden not measured variables



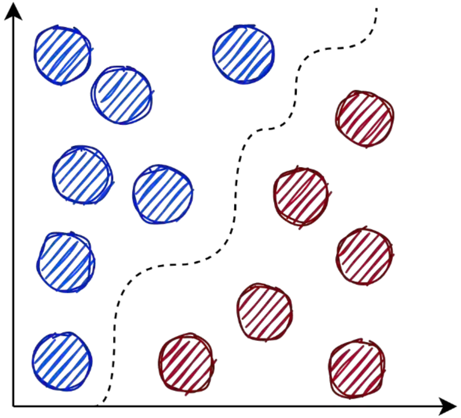
Original Data

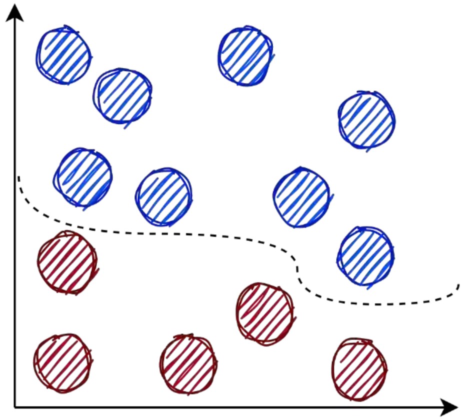
Concept Drift Problem

Formal Description:

➔ Concept drift: a phenomenon in which the statistical properties of a target variable change over time in an arbitrary way

➔ Reasons: Changes in hidden not measured variables

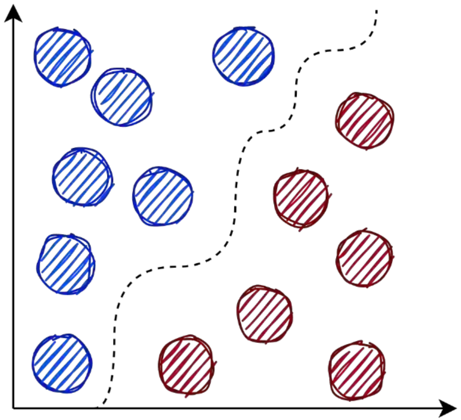

Original Data


p(Y|X) changes

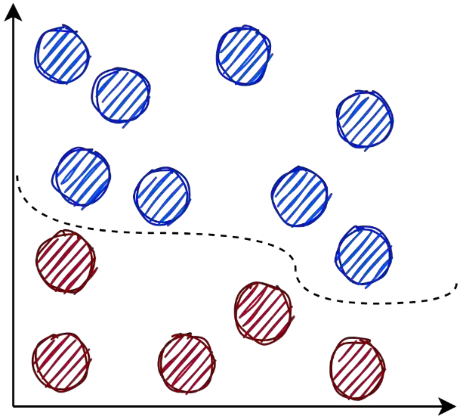
Concept Drift Problem

Formal Description:

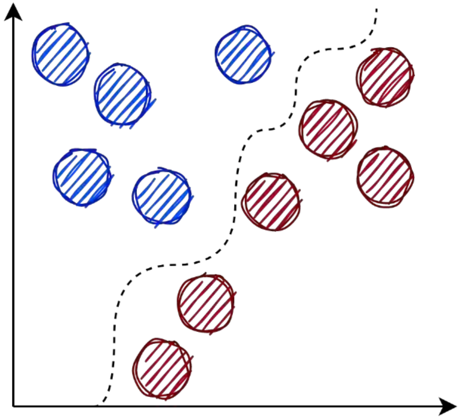
- ➔ Concept drift: a phenomenon in which the statistical properties of a target variable change over time in an arbitrary way
- ➔ Reasons: Changes in hidden not measured variables



Original Data



p(Y|X) changes



p(X) changes

Research Goals

- *Research Goal I: We aim to develop a new Data management solution suitable to IoT environments requirements. By combining the main approaches of **active**, **actor-oriented** and **array databases**, we aim to provide a scalable, reactive and efficient data management and analysis system for IoT.*

- *Research goal II: We aim to integrate **ML inference** into the data management system, enabling optimizations that would not be possible when treating ML analytics as independent processes.*

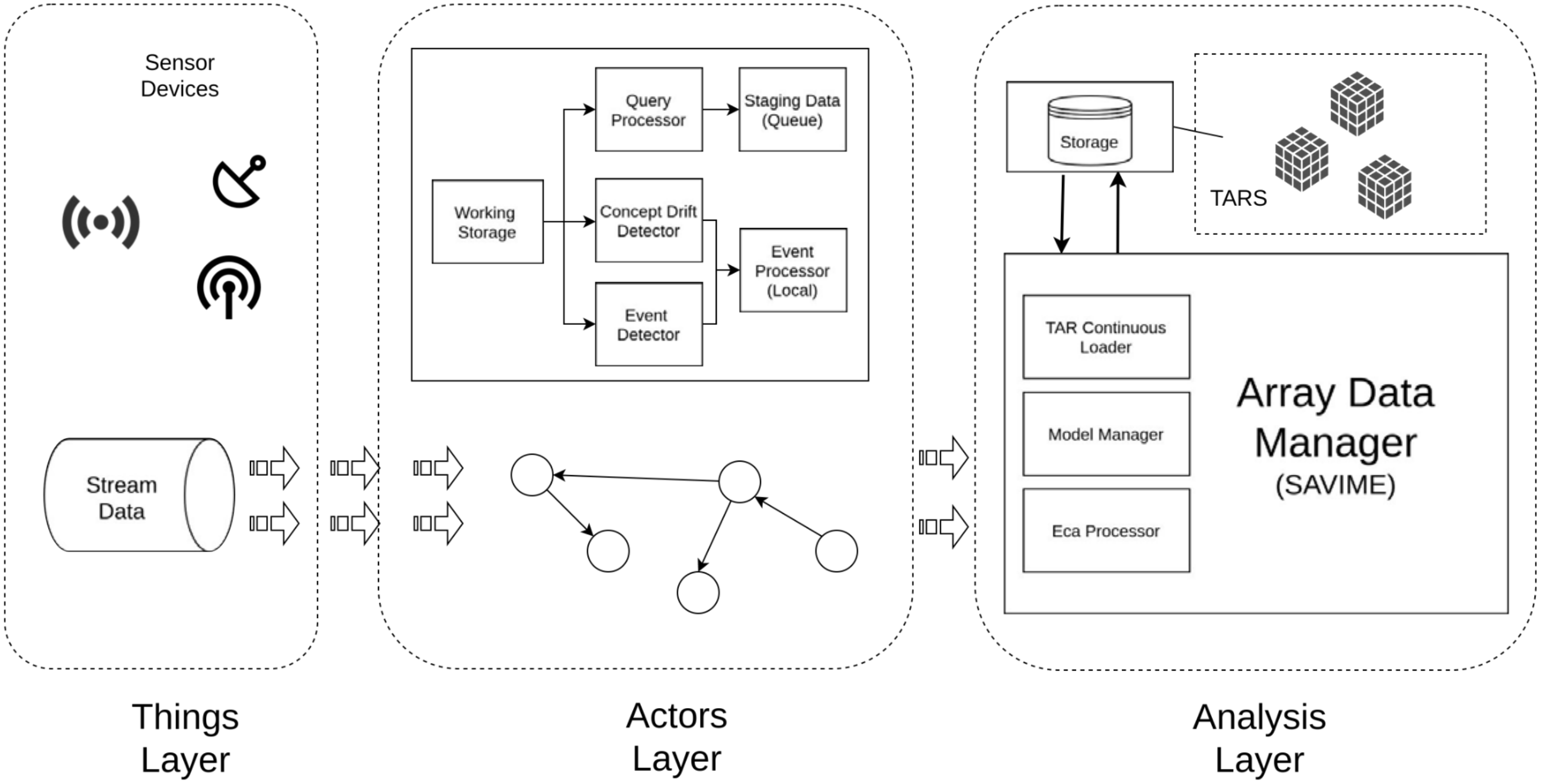
- *Research Goal III: We aim to integrate an adaptive **concept drift** detection method into our solution that considers the characteristics of IoT data.*

Research Goals

System Features		Actor Oriented Databases	Array Databases	Active Databases	Proposed Solution
Actor-Based Programming	Dynamic Scalability Asynchronous primitives Encapsulation	+	-	-	+
Array Based Data Management	Array-Based Operations Flexible Storage Format	-	+	-	+
Complex Event Handling	Event Detection Reactive Behavior	-	-	+	+
Machine Learning Support	ML as first class operations Concept Drift Handling	-	-	-	+

Potential contributions from different models for IoT data management

Solution Design



Future Directions

- ➔ Real Scenarios: The solution should be evaluated on a weather prediction and monitoring scenario
- ➔ Comparative Experiments: Compare the solution to an approach that integrates different already existing data systems
- ➔ Metrics: Measure scalability and performance of the solution when compared to state-of-the-art techniques and systems

Future Directions

- Real Scenarios: The solution should be evaluated on a weather prediction and monitoring scenario

- Comparative Experiments: Compare the solution to an approach that integrates different already existing data systems

- Metrics: Measure scalability and performance of the solution when compared to state-of-the-art techniques and systems

Thank you for your interest