University of Padua

Department of Information Engineering

Master Degree in ICT for Internet and Multimedia

Leveraging cloud and machine learning technologies for the development of a knowledge IOT database

Master thesis

Relator

Prof. Lorenzo Vangelista

Master Candidate

Alessandro Discalzi

ID 2088235

Academic Year 2023-2024

Summary

This document describes the work done during the 750-hours final project at 221e S.r.l.
The project's goal is to architect and develop a cloud-based system capable of ingesting and processing data from heterogeneous IoT sensors so that a knowledge database can be built.
The system must be designed to be scalable and fault-tolerant, and it must be platform-agnostic.
This document is going to describe the company, the idea behind the project, the work done and an assessment of what I developed and learned during my internship.

"If the past is just dust
Then the future could be our dream"

— Lorna Shore

Acknowledgements

Prof. Lorenzo Vangelista, my thesis supervisor, deserves my deepest gratitude for his exceptional support and guidance throughout the completion of this research.
My family, for their encouragement and understanding throughout this academic endeavour, has my heartfelt thanks.
I am truly grateful to Luca Perosa, Bledar Gogaj, Marco Lionello, and all my peers at SCAI ITEC, for their unwavering support when I made the decision to pursue a Master's degree.
I extend my sincere appreciation to PhD. Roberto Bortoletto, my company tutor, and all my colleagues in 221e for their invaluable support and guidance throughout my final project.
Last but not least, I want to give a shoutout to all my friends for having my back and just being there through thick and thin. Your friendship means a lot to me, and I appreciate the support and good times we've shared.

Padova, October 2024 Alessandro Discalzi

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Company

221e S.r.l. is an Italian company based in Abano Terme (PD). The company, founded in 2012, is a leading supplier of IoT solution providing a wide range of products, both hardware and software, for industrial and wereable applications. The great experience and know how of the company's team, allows to provide clients with the best hardware solution for their needs, which can be enhached via the company's software platform or via a custom solution.

Figure 1.1: 221e's logo

1.2 Idea

The idea behind the project is to create a new cloud-based software platform for the company's IoT devices. The platform needs to ingest data from a variety of IoT devices, process it and create a knowledge database that can be used to provide insights and predictions to the end user.

1.3 Thesis outline

The second chapter describes the high level requirements.

The fourth chapter describes how the solution has been implemented.

The fifth chapter assess the results of the project.

The last chapter describes the conclusion of the project and the future work.

Chapter 2

Objectives and requirements

2.1 Data collection

2.2 Security

2.3 Cloud infrastructure

2.4 Scalability

Chapter 3

Technologies

In this chapter it's reported the study made on the various technologies taken into account to develop the project.

3.1 Cloud

This section describes the services offered by Amazon Web Services and Microsoft Azure, exploring their features, pros and cons. These services are the most used cloud services in the world and they offer a wide range of services that can be used to develop the project. It's also important to mention that these two providers were chosen right away because of the already developed experience with them, both in the company and in the author of this thesis.

3.1.1 Amazon Web Services

Batch

AWS Batch is a fully managed service that enables developers to easily and efficiently run thousands of batch and machine learning computing jobs on AWS.
Pros

- Fully managed
- Scalable
- Cost effective
- Supports different batch processing scenarios
- Supports machine learning
- Easy to use
- Versatile

Cons

- Not well documented

DynamoDB

AWS DynamoDB is a fully managed NoSQL database service that provides fast and predictable performance with seamless scalability. Tables can store and retrieve virtually any amount of data, serving any level of request traffic. It automatically spreads the data and traffic for the table over a sufficient number of servers to handle the request capacity specified by the customer and the amount of data stored, while maintaining consistent and fast performance.
Pros

- Fully managed
- Fast and predictable performance
- Scalable
- Highly available
- NoSQL

Cons

- Hard to make changes against bulks of records
- Need to know at prior which queries will be made

Elastic map reduce (EMR)

AWS EMR is a big data platform that simplifies the deployment and management of big data frameworks, like Apache Hadoop and Apache Spark, on AWS.
Pros

- Fully managed
- Scalable
- Petabyte scale data processing
- Easy resources provisioning
- Reconfigurable

Cons

- Complexity
- Costly

Glue

AWS Glue is a fully managed ETL service that enables efficient data integration on a large scale.
Pros

- Fully managed
- Pay per use

- Scalable
- Provides a centralize metadata repository
- Supports different data sources and formats
- Can automatically discover and catalog data from various sources
- Allow for job scheduling
- Data encryption

Cons

- Costly for high workloads
- Performance issues with large datasets
- Complexity

Greengrass

AWS Greengrass is an open source edge runtime and cloud service used to build, deploy, and manage device software. It enables the devices to process the data locally, while still using the cloud for management, analytics, and durable storage.
It also enables encryption at rest and in transit and it can also extend device functionality with AWS Lambda functions.
Pros

- Edge computing
- Encryption at rest and in transit
- Extend device functionality with AWS Lambda functions
- ML models deployment

Cons

- Restrained to AWS services
- Not platform agnostic
- Resource intensive for small devices
- Need a connection for the initial setup

IoT Core

AWS IoT Core is a fully managed cloud service that lets connected devices easily and securely interact with cloud applications and other devices. It is composed of multiple services like Device Management, Device Defender, Device Advisor, and IoT Analytics and only some of them can be used during the development.
Pros

- Composed of multiple services so only the necessary ones can be used
- Encription at rest and in transit

- Supports MQTT, HTTP, and WebSockets
- Allows for device management
- Allows for machine learning at edge
- Can trigger events thanks to custom rules

Cons

- Not platform agnostic if installed on devices
- Lacks of integration for some devices

Kendra

AWS Kendra is a fully managed enterprise search service that allows developers to add search capabilities across various content repositories leveraging on built in connectors.
Pros

- Fully managed
- Scalable
- Supports multiple data sources
- Easy to use and set up
- Accurate search results

Cons

- Costly

Kinesis Data Firehose

AWS Kinesis Data Firehose is a fully managed service that simplifies the process of capturing, transforming and loading streaming data. It acts as an ETL service that can capture, transform, and load streaming data into a variety of AWS services. Additionaly it can transform raw data in column oriented data formats like Apache Parquet
Pros

- Fully managed
- Can read data from IoT core and Kinesis Data Streams
- Scalable
- Can transform data
- Can load data into different AWS services
- Supports batching based on time or size

Cons

- Not always cost effective
- Limited transformation capabilities
- Does not support batching based on more complex rules

Kinesis Data Streams

AWS Kinesis Data Stream is a fully managed service that simplify the capture, processing and loading of streaming data in real time at any scale thus enabling real-time data analytics with ease.
Pros

- Fully managed
- Scalable
- Real-time and fast data processing
- Keeps data for 24 hours by default

Cons

- Not always cost effective
- Limited data retention
- Limited data transformation
- Not useful for certain batch processing scenarios

Lake Formation

AWS Lake Formation is a fully managed service that simplifies the creation, security and management of data lakes. It allows for cleaning and transforming the data using machine learning.
Pros

- Fully managed
- Scalable
- Secure
- Simplifies lake creation
- Simplifies ingestion management
- Simplifies permission management
- Provides data auditing
- Supports machine learning
- Supports data cataloging

Cons

- Complexity
- Costly
- Not native support for all data sources

Lambda

AWS Lambda is an event driven serverless compute service that automatically manages the underlying compute resources. AWS Lambda can be used to extend other AWS services with custom logic, and to create new back-end services that can operate at AWS scale, performance, and security.
Pros

- Fully managed
- Serverless
- Pay per use
- Scalable
- Easy to integrate with other AWS services
- Supports multiple programming languages
- Easy to deploy and maintain
- Can run parallel executions
- Low time to market
- Supports custom libraries

Cons

- Limited execution time (15 mins)
- Limited memory
- Limited environment variables
- Maximum 1000 concurrent executions
- Cold start
- Not cost effective for high workloads

Managed Service for Apache Flink (MSF)

AWS MSF is a fully managed service that simplifies the creation and the execution of real time application using Apache Flink .
Pros

- Fully managed
- Scalable
- Supports batch and stream processing
- Real-time processing
- Large-scale data processing

Cons

- Complexity

Managed Streaming for Kafka (MSK)

AWS MSK is a fully managed service that simplifies the setup, the scaling and the management of Apache Kafka clusters.
Pros

- Fully managed
- Scalable
- Cost effective
- Secure
- High availability
- Easy to integrate with other AWS services

Cons

- Local testing challenges: hard to replicate the same environment in production and locally
- Not suitable for high traffic scenarios
- Complexity

Sage Maker

AWS Sage maker is a cloud-based machine learning platform that allows developer to build, train and deploy machine learning models.
Pros

- Fully managed
- Scalable
- Supports multiple machine learning frameworks
- Supports multiple programming languages
- Allow for easy model deployment

Cons

- Cannot schedule training jobs
- Costly for high workloads

Simple Storage Service (S3)

AWS S3 is an object storage service offering scalability, data availability, security, and performance. With S3, any amount of data can be stored and retrieved from anywhere on the web. Data is stored as objects in buckets, with each object representing a file and its metadata.
Pros

- Scalable

- Highly available
- Secure
- Durable
- Cost effective
- No bucket size limit
- No limit to the number of objects that can be stored in a bucket
- Has different storage classes to fit frequent access, infrequent access, and long-term storage

Cons

- Not suitable for small files
- Object size limit (5TB)
- Maximum 100 buckets per account
- Max 5GB per file upload via PUT operation

3.1.2 Microsoft Azure

3.2 Present Solutions

3.2.1 Eclipse Kura

3.2.2 Eurotech Everyware Cloud

3.2.3 STMicrelectronics

3.3 Machine Learning at the Edge and Federated Learning

3.3.1 Tensorflow Lite

3.3.2 Tiny Engine

Chapter 4

Methodology

Introduction

4.1 Data Collection

4.2 Architecture

Chapter 5

Results

Chapter intro

5.1 Tests

Chapter 6

Conclusion

6.1 Objectives achieved

6.2 Future developments

6.3 What I learned

6.4 Final considerations

Chapter 7

Bibliography

Bibliographic references

James P. Womack, Daniel T. Jones. *Lean Thinking, Second Editon.* Simon & Schuster, Inc., 2010.

Website references

Manifesto Agile. URL: http://agilemanifesto.org/iso/it/.