1. Technical Specification Double-side Page

2. **TECHNICAL SCOPE:** Summarize the solution developed during the EXPERIMENT phase: how have you finally addressed the challenge/Theme Challenges and tackled with its requirements and data. Include a diagram.

FreshForecast is a Big Data solution that allows (i) massive and continuous data ingestion and curation pipelines in order to blend the different data sources , and (ii) supports different ML pipelines for continuous modeling and prediction services.
From a biz perspective, FreshForecast is an ordering decision support system to reduce food waste by providing accurate sales prediction trends and an extended view of the inventory.
During the EXPERIMENT phase, we downscaled our development cluster to fit better the planned MVP. Specifically, we rely on 3 nodes with a total of 48 vCPU and 192 GB RAM. The cluster handled the sample data and modeling scenarios with good performance and it looks reliable for future larger data sets and computing routines. We designed the solution for batch processing, allowing for updated sales forecasts at regular intervals, such as twice per day, of course, depending on more frequent data sources. Regarding the continuous modeling process, FreshForecast provides daily new models. However, new models need to be analyzed and promoted by Sales Analysts to be deployed.
FreshForecast proposes a multilayer data structure for all data life cycles. At a starting point, an ingestion layer is populated from many data providers through different media. Following, comes the curation layer in which ingested data are validated, transformed and blended. Finally, it comes the provisioning layer populated by intuitive multidimensional structures describing (i) sales history, such as weather conditions during harvest or along the supply chain, (ii) sales models for product types and categories as well as different types of stores (small/large), and (iii) forecast results. From this data layer are tackled most of the information requirements from the biz and technical side.
FreshForecast modeling approach is based on a Deep Learning (DL) architecture. During the Experiment phase, it had evolved from the original proposal to fit provided sample data profile. Basically, to adapt to sparse sale scenarios and some limitations on logistic data (mainly ordering data). At a higher level, it combines a Multilayer Perceptron (MLP) classifier that predicts sales trends for the next 3 days and contextualizes with relevant data in order to support the ordering decision making.
See architecture diagram for more details on Annex 2.
During EVLOVE phase, we ported the developed solution onto a managed databrick cluster as our production evironement.
We also connected the module to your existing FreshAnalytics Management platform, thus enabling OAuth 2.0 and multi user / multi tennant administration through keycloak

3. **ALGORITHMS, TOOLS AND CONCLUSIONS:** Detail the algorithms and tools finally selected to accomplish the challenge/Theme Challenges. Summarize the main results that you have obtained during the EXPERIMENT phase: data, insights, conclusions and the main contributions to solve the challenge/Theme Challenges.

FreshForecast is based on a Spark cluster, a fast and easy to use cluster computing engine. Scala is the native programming language. It is a language that gives easy support for functional programming. Therefore, having the lock-free data processing of functional programming and an easy parallel processing, the build of required reliable and scalable data pipelines is straightforward.
Regarding data service, FreshForecast implements data storage and access by using Hadoop and Hive service. Hadoop gives scalable and interoperable storage capabilities and Hive gives data warehousing capabilities on top. For data ingestion, FreshForecast dealt with different batch data deliverables and it was agreed with DP that a FTP-like strategy could be followed to receive them.
FreshForecast DL architecture combines a MLPC suitable for classifying the next 3 days sales trends and contextual data that complete the predicted scenario. Specifically, FreshForecast provides together the predicted sales trends for next 3 days and the current status of the inventory. Not only considering the quantity, but also last waste quantities, weather at last-mile transport and storage
The DL architecture is implemented using MLlib, Spark's machine learning (ML) library. It provides us tools to make practical ML scalable and easy. Specifical tools are, the deep feedforward NN for regression algorithm, feature transformations, ML Pipeline construction, model evaluation, and saving and loading models and pipelines. Forecasting accuracy is determined through cross validation and sliding window validation.
FreshForecast implements analysis and data exploration capabilities by using Apache Superset, an intuitive interface for visualizing datasets and building interactive dashboards. Different roles and related information requirements are considered, such as the Warehouse Buyer and the Store Manager. For deeper exploration of the data, required by tech roles, such as DevOps or Data Scientist, FreshForecast implements an interactive notebooks system using Apache Zeppelin.
FreshForecast also implements a basic prediction service. Specifically, it was based on a REST API implemented with Akka. It was conceived for two reasons: (i) to explore the possibility of prediction as a service biz strategy, and (ii) in the context of the collaboration with CYC.
During the EVOLVE phase we were able to mature the models, to quantify the model performance and provide a coefficient of determination (r^2) as well as the mean square error during cross validation.

4. **SCALABILITY AND FLEXIBILITY OF THE SOLUTION:** Explain how the solution copes with the challenge/Theme Challenges requirements and how it can be adapted to other similar problems. What work is still pending to create a real/stable product if any? What TRL level is it in?

At operational level, FreshForecast overcomes the computing efforts arising from humongous and increasing datasets as well intensive ML routines by relying on one of the best proven Big Data stacks: Hadoop, Hive and Spark. A tech stack natively distributed, reliable, scalable and easy to use. FreshForecast can straightforward scale through the number of hosts and the resources on them . But also, hosts could run on the cloud or on-premise.

FreshForecast uses Apache project Ambari that makes out of the box the provisioning, managing, and monitoring of Hadoop clusters. However, we have tested some of our pipelines on Databricks, a Big Data services leading provider. We found that relying on this infrastructure provider we could reduce the team size and the feasibility of an immediate go live process. On this matter, we have met Databricks and AWS consultants and the pricing expectations have been promising.

On the other side, to overcome the diverse data providers scenario, FreshForecast relies on Scala. A programming language easy to use and with less code. Therefore, FreshForecast can build flexible and compact pipelines that could be easy to adapt and include further data sources. Indeed, during the last milestone we have integrated new pipelines to consider last arrived data describing stores and orders.

FreshForecast can also overcome other related domains. Main reason is our development process following a Model Driven Architecture (MDA). Mainly, considering information requirements and needs at early stages of the design. Therefore, it is feasible to adapt later stages, such as the conceptual design and the implementation.

5. **DATA GOVERNANCE AND LEGAL COMPLIANCE:** Describe the security level of the solution, i.e. how authentication, authorization policies, encryption or other approaches are used to keep data secure. Explain how the solution is compliant with the current data legislations concerning security and privacy (e.g. GDPR).

Legal perspective: The development of the solution uses a privacy-by-design approach. FreshForecast is compliant with EU data privacy legislation as well as the upcoming Data Governance Act, expected to be released this summer. tsenso has developed a consistent liability concept for food safety data generated by the system.

Operational: Regarding Data sharing and governance, at a starting point, FreshForecast manages the casuistic of different user roles on Biz and technical side, and multiple data providers (sometimes competitors between them). Specifically, from the Biz side, FreshForecast considered the Warehouse Buyer and the Store Manager, and from the Tech side, the Data Scientist and DevOps.

In order to implement the strategy, FreshForecast relies on three axes. First, the pseudo anonymization process planned for the early stages of the data in order to protect critical total and subtotal values. Then, the data containers coming from the multilayered data structure and DataMarts. Finally, an authentication protocol to prove identities in a secure manner that depends on deployed Big Data infrastructure. In case a solution is deployed on-premise at MIGROS, it can be kept behind the corporate firewall, adding another layer of security.

Given the increasing number of data encryption and poisoning attacks, tsenso is currently active in a research project https://poison ivy.de/to detect and prevent data & model poisoning.

6. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Describe the quality process followed for the final product. Technologically, which problems have you encountered and how you have solved them, and any processes followed that guarantee that the solution fulfills the challenge/Theme Challenges and data provider requirements.

The accuracy and reliability of any forecasting solution can only be as good as the data it is built on. At tsenso, we follow a strict and scientific approach to the design of the solution architecture as well as the respective models. Despite focusing on AI methods, we still believe in the power of deterministic analytics descriptions and use such to validate our results.

FreshForecast accomplishes QA at different points of the data lifecycle, ensuring the validity and correct operation of the different data pipelines. At data ingestion, where digital data agreements are used to accept the data delivery. Then, at data curation, where data agreements are used for data validation. Finally, at provisioning level, where the accuracy of forecasts and models are monitored, and product histories and details are validated. In the different QA processes, different user roles take the responsibility. Specifically, FreshForecast supports DevOps and Data Scientist users for those tasks. DevOps role is responsible for product related tasks and data scientist for the forecast related ones.

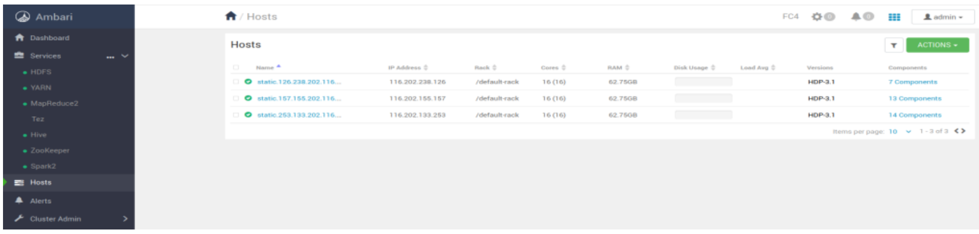
During operation, one of the key challenges of demand forecasting solutions is to assure that the models used stay accurate and do not diverge. We are currently working on a module for automated model calibration, based on the principles of uncertainty quantification, allowing for a regular and automated quality check and updating the models.

FreshForecast early identified a particular potential risk at the development and testing phase of the project and later one it had been suffered. That is to deal with the accuracy of Forecasts and Models. Specifically, the use of RNN was discouraged mainly for the sparse attribute of sample data and the need of ad-hoc ml models to overcome it. However, some promising experiments had been performed using DeepLearning library and the recurrent layers available. To overcome the problem in this phase and complete the DVC, we focus on obtaining trends for next days and considering a window of previous sales. In this way, the grain of the complete scenario diffused the sparse nature of it.

8. Annex 1. Means for accessing the MVP

Please, indicate in 1 page indicating the means for accessing the MVP for a potential customer (login information, website address, link to a demo video or whatever means are needed to check that the MVP exists and works).

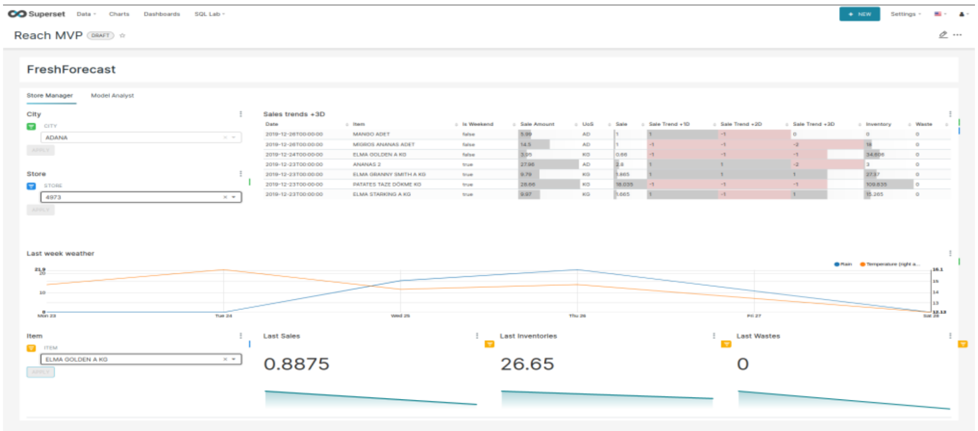
We have developed a dashboard reproducing the Store Manager UI proposed in previous mockups to support the ordering decision to reduce waste.
Specifically, UIs are provided on the Superset server (116.203.76.216:8080) populated directly from the Provisioning layer (Hive) allocated in the development cluster (Ambari). For security reasons, Superset ports are closed and only accessible through SSL tunnel. However, do not hesitate in asking for access if it is required. In any case we will have access from our machines during the presentation. Following, you have snapshots of Cluster Manager and some Dashboard instances.



Hosts

	Name	IP Address	Rack	Cores	RAM	Disk Usage	Load Avg	Versions	Components
□	static.126.238.202.116...	116.202.238.126	/default-rack	16 (16)	62.75GB			HDP-3.1	7 Components
□	static.157.155.202.116...	116.202.155.157	/default-rack	16 (16)	62.75GB			HDP-3.1	13 Components
□	static.253.133.202.116...	116.202.133.253	/default-rack	16 (16)	62.75GB			HDP-3.1	14 Components

Items per page: 10 1 - 3 of 3



Sales trends +3D

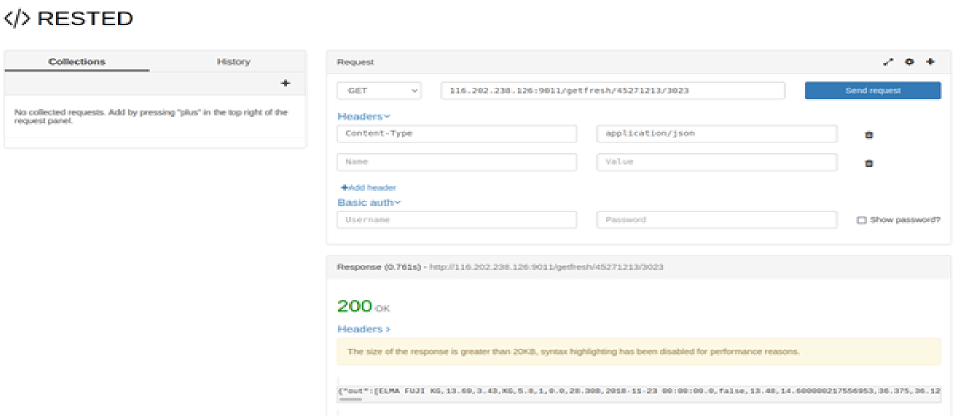
Date	Item	Is Weekend	Sale Amount	UoS	Sale	Sale Trend +1D	Sale Trend +2D	Sale Trend +3D	Inventory	Waste
2019-12-26T00:00:00	MANGO ADET	false	5.99	AD	1	1	-1	0	0	0
2019-12-26T00:00:00	MIGROS ANANAS ADET	false	14.5	AD	1	-1	-1	-2	18	0
2019-12-24T00:00:00	ELMA GOLDEN A KG	false	3.95	KG	0.66	-1	-1	-1	34.606	0
2019-12-23T00:00:00	ANANAS 2	true	27.96	AD	2.8	1	1	-2	3	0
2019-12-23T00:00:00	ELMA GRANNY SMITH A KG	true	9.79	KG	1.865	1	1	1	27.37	0
2019-12-23T00:00:00	PATATES TAZE DÖKME KG	true	28.66	KG	18.035	-1	-1	-1	109.835	0
2019-12-23T00:00:00	ELMA STARKING A KG	true	9.97	KG	1.665	1	-1	1	15.265	0

Last Sales
0.8875

Last Inventories
26.65

Last Wastes
0

The MVP of our prediction service can be accessed by API:



No collected requests. Add by pressing "plus" in the top right of the request panel.

GET 116.202.238.126:9011/getfresh/45271213/3023

Content-Type: application/json

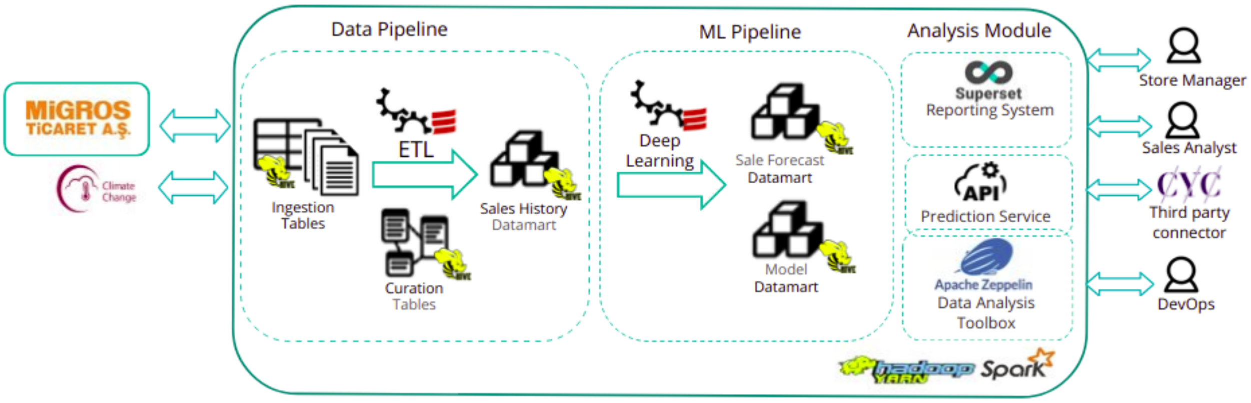
Response (0.761s) - http://116.202.238.126:9011/getfresh/45271213/3023

200 OK

The size of the response is greater than 20KB, syntax highlighting has been disabled for performance reasons.

```
{"out":[ELMA FUJI KG,13.69,3.43,KG,5.0,1,0.0,28.308,2018-11-23 00:00:00.0,false,13.48,14.600000217556953,36.375,36.12
```

9. Annex 2. MVP architecture diagram


MIGROS TİCARET A.Ş.
Climate Change
Data Pipeline
Ingestion Tables
ETL
Sales History Datamart
Curation Tables
ML Pipeline
Deep Learning
Sale Forecast Datamart
Model Datamart
Analysis Module
Superset Reporting System
API
Prediction Service
Apache Zeppelin Data Analysis Toolbox
hadoop YARN
Spark
Store Manager
Sales Analyst
CYC
Third party connector
DevOps