UST

Azure AI and advance analytics capabilities

Accelerate your AI journey

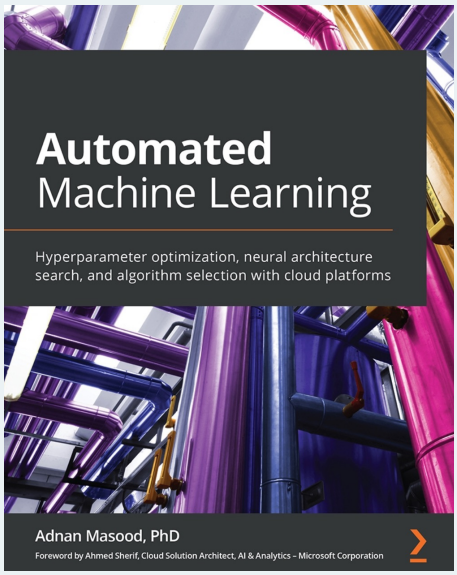
ust.com

Microsoft Azure is a market leader in advance analytics – building upon how AI and machine learning are reinventing the digital transformation market. To maximize your business outcomes, your organization needs the right expertise in Azure AI eco-system, where UST serves as the trusted partner. With deep expertise in Microsoft Azure Bot Framework, Azure machine learning and cognitive services, UST accelerates your AI journey and extend operational & enterprise systems to fulfill the automation promise.

UST provides Azure specific expertise in the following areas:

- AI strategic roadmap and playbook development
- Machine learning solutions development & deployment
- Visualization, data-driven insights and analytics
- Fairlearn for ethical, responsible & interoperable AI

With Azure AI MVPs and regional directors, following are some of the recent publications UST has collaborated with Microsoft and Packt.



Automated Machine Learning by Dr. Adnan Masood – Chief AI Architect @ UST and Microsoft Regional Director

Microsoft Practice AI Development Playbook – UST Contributing Authorship

UST's seasoned data scientists, AI engineers, and machine learning developers equip your organization with best practices, and helps you get up to speed with the latest Microsoft Azure machine learning capabilities - to build, train, and deploy machine learning models quickly and easily. Their expertise help you utilize cutting-edge technologies like automated machine learning, Fairlearn, Jupyter, Visual Studio Code frameworks such as PyTorch Enterprise, TensorFlow, and Scikit-learn. With low code and no code tools like automated machine learning and drag and drop interface, you can expand your data science teams and generate models faster. In addition, using ONNX Runtime, you can simply deploy at scale using Azure Kubernetes Service (AKS) and maximize machine learning inference.

UST has implemented AI solutions across a range of real time use cases like recommendation engines, predictive maintenance, fraud detection, and predictive models. You can see some of the case studies below.

Data ingestion and migration

For a major healthcare provider based in Pennsylvania

Problem

Our client wanted to migrate data from existing IBM Netezza on-prem to Azure cloud.

Solution

- Set up Azure Cloud environment
- Creation of data pipelines in Azure Data Factory for data ingestion from:
  - IBM Netezza on-prem
  - Epic environment (Clarity database)
- Storage of data in ADLS Gen2
- Azure Synapse Analytics service to manage and serve data for reporting
- QlikView is used to create operational and analytical reports

Impact

- 40% increase in efficiency (from 6-7 hours to 3-4 hours to complete the daily load)
- High scalability
- Number of users: up to 30K
- Number of tables migrated from Clarity to Synapse: 1600
- Daily data pull: 300 – 400 GB
- Historical data migrated from Netezza to Synapse: 15 – 16 TB

Tech stack

Data intelligence management and analytics

For a Fortune 500 FMCG company

Problem

Manage, connect, and leverage R&D data for better insights.

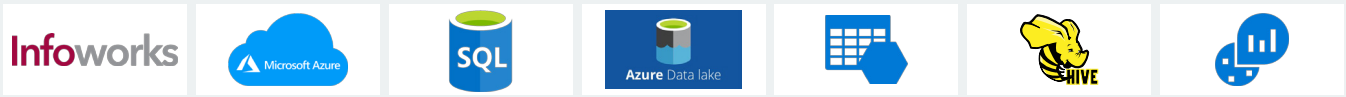
Solution

- Portal for R&D, data scientists, product developers, and other users to search for projects/ products based on keywords(fuzzy match)
- Allow users to select project/product IDs, add selected items to a cart, check out the cart and prepare related files
- Write the files to secure Blob location and notify the user about the Blob location

Impact

- Search and prepare relevant data based on project/ product ID
- Secure Blob location as data repository to derive insights
- Reduced time to analyze data by ~75%
- Reduced admin and infrastructure cost by ~50%
- Number of rows/columns: 900k+

Tech stack



Multi-source data ingestion
For a Fortune 500 CPG company

Problem

Our client wanted to create a unified data lake to enable data science team to leverage data to derive insights.

Solution

Cloud based solution to ingest data from multiple sources to a managed data lake.

- Total table count: 300+
- Data source: Oracle, Teradata, DB2
- Refresh frequency: hourly, daily, weekly, custom time
- Load type: Incremental load/truncate and load

Impact

- Managed data lake as a repository for both structured and unstructured data
- Minimize cost and give insights for end users

Tech stack

Data lake creation and reporting services
For a global retail conglomerate's US arm

Problem

Our client wanted to support their reporting, analytical and archival needs in a unified data lake and replace the outdated reporting platform.

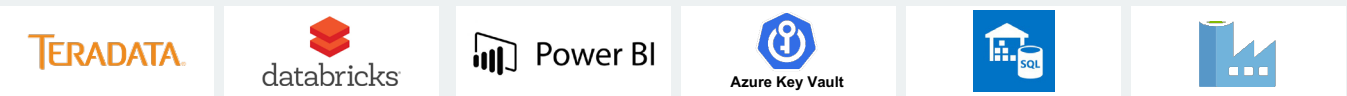
Solution

- Azure Data Factory to ingest data from disparate sources
- Transformations and data-driven workflow for orchestrating and automating data movement through a combined pipeline with an execution schedule
- Automatic splitting of workloads across various processors which scales up and down on demand
- Data lake in Azure Data Lake Storage
- Data warehouse to store all final transformed data with Power BI connected to it for report generation

Impact

- Support a user base of 300-400 individuals
- Approximate monthly data volume of 2.5TB
- Data size handled: Cyborg (~55GB/day), SF (~10MB/day), PeopleSoft (~6GB/day), Kronos (~5MB/day), UtliPro (~20GB)
- Increase efficiency and effectiveness of reporting

Tech stack



Data ingestion and transformation

For a global retail conglomerate's US arm

Problem

Our client wanted to fetch data from existing multivariate sources, perform transformations, and write it back to destination storage , orchestrating and designing data pipelines.

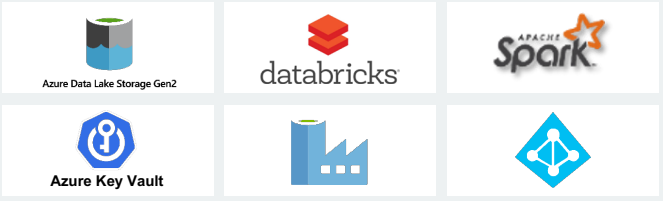
Solution

- Data sources – Mainframe files (VSAM, SEQ), DB2, Informix DB, SQL Server & Oracle
- ETL from data source using Azure Databricks
- Databricks and HDInsights on Azure Cloud
- Storage of source data, intermediate and destination data in ADLS Gen2
- Creation of data pipelines in Azure Data Factory(ADF)
- Failure e-mail notification using Logic Apps
- Azure DevOps as code repository and for versioning
- HDFS using Hive for storing source tables, testing and debugging

Impact

- Output files (CSV) on runtime (either once or multiple times a day)
- Data size handled: 25+ GB/day

Tech stack



Data integration

For a global retail conglomerate's US arm

Problem

Our client wanted to make the required data available in ADLS and respective vendor servers for the forecasting teams.

Solution

- IBM datastage to transfer the files from Mainframes to data lake via Azure Virtual Machines (VMs)
- SFTP wrappers to pull/push the data files from external vendor servers via Azure VMs
- CLI utility wrappers to upload/download the data files from cloud storages such as Amazon S3 via Azure VMs
- Azure Data Factory pipelines to move the data from Azure VMs to ADLS Gen1 data lake storage
- In the case of ADLS Gen2 data lake, the containers are mounted into the Azure VMs, and the data is copied from/to data lake directly by the SFTP and CLI wrappers.
    - The Azure Data Factory pipelines have been eliminated

Impact

- Data size handled: ~10 GB approx daily
- Ingestion type: Batch
- Data format: txt, CSV, mainframe flat files
- Ingestion speed/frequency: every 30 mins, every 1 hour, daily, weekly, monthly and quarterly

Tech stack

Data migration (on-premise to cloud)

For US-based shopping channel and e-commerce retailer

Problem

Ingest and transform data for various markets from on-prem to Azure Cloud.

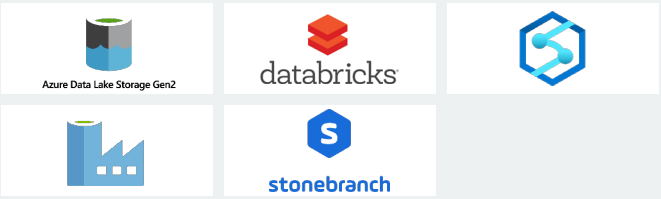
Solution

- Proposed a metadata driven processing framework which is developed using PySpark
- The framework dynamically handles multiple source files coming from various markets by fetching the configuration values from underlying metadata tables.
- Once the data is landed in ADLS landing zone, it then goes through multiple processing layers like rule based cleansing, deduplication and transformations
- The processed data is then loaded into persistent storage in Delta Lake and Azure Synapse target tables.
- The process is orchestrated with the help of Stonebranch and Azure Data Factory.

Impact

- Minimal configuration required for processing different files from various markets
- Parallelism can be implemented for all independent processes whereas the dependent ones are run in the appropriate order as per our configurations
- Stonebranch and ADF enable scheduling and automation of pipeline runs

Tech stack



End-to-end data virtualization using Denodo

For a multinational F&B corporation, headquartered in New York

Problem

Our client wanted to create a global dashboard to track and manage waste of all markets under one umbrella. The client had no centralized reporting and all the business units were managing their waste independently.

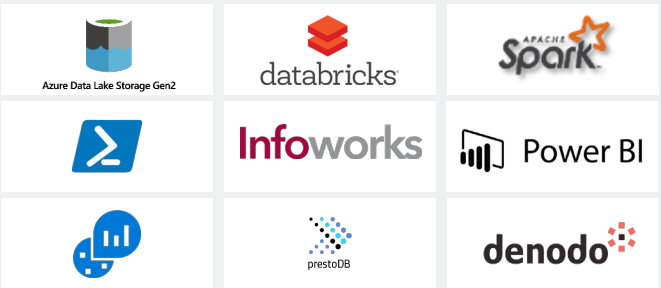
Solution

- Developed an automated process that collects XLS files uploaded by the business users and refreshes the dashboards
- Used Denodo to:
  - create views and store business logic
  - federate data of warehouse (Teradata) and XLS (in ADLS, accessed through Presto) to create business insights
- Used PowerBI as the visualization tool
- Used a scheduler for automatic refresh

Impact

- Data size handled: ~5 GB
- Ingestion type: File Watcher
- Ingestion speed/frequency: ad-hoc, based on business file arrival

Tech stack



Geospatial data ingestion
For a global FMCG giant

Problem

Our client uses Geotab, a global positioning system. Communicating with Geotab database through API has limitation of 50000 records per API call. The challenge is to analyze partial results received through each API call.

Solution

- Cloud based solution to ingest the data from Geotab databases to a managed data lake
- Data-pipelines ingest data on a weekly basis for incremental data
- Slowly Changing Dimensions (SCD) handling

Impact

- Managed data lake to provide complete dataset for analysis

Tech stack

Big data administration
For a major F&B corporation

Problem

- Manage downstream analytics
- Support maintenance activities for the data lake

Solution

- Data lake maintenance with real-time integration to marketing, shipping, logistics, sales and customer data
- Apache Nifi is being used for data routing and transformation, Infoworks for data orchestration and Trifacta for data wrangling
- 4 data bricks workspace managed
- 13 HDI clusters are managed, out of which 6 are with Enterprise Security Package and integrated with Active Directory Kerberos and Ranger
- 6 ADLs managed

Impact

- Reduced admin and infrastructure cost by more than 50%
- Number of users: 1000+
- Data volume (monthly): 8 - 10 TB

Tech stack

Hospital readmissions analytics

For a large healthcare payer based in Louisiana

Problem

Identify patients with high probability of (unplanned) readmission within 30 days of discharge

Solution

- Data fetch and pre-processing: Raw data is fetched from server and pre-processing pipeline is run
- Feature engineering: Longitudinal features are built per patient summarizing his/her timeline
- Readmission prediction model development and validation: User input is taken for the attributes like model type to run, hyper parameter variables and their values, type of cross validation strategy, validation period, etc.
- Model evaluation: Model metrics report generated that includes AUC-ROC, AUC-PR, Precision, Recall, Precision @ N% recall etc.

Impact

- Reduced readmission rates
- Enabled targeting of high-risk members with personalized interventions

Tech stack

For more than 20 years, UST has worked side by side with the world's best companies to make a real impact through transformation. Powered by technology, inspired by people and led by our purpose, we partner with our clients from designto operation. Through our nimble approach, we identify their core challenges, and craft disruptive solutions that bring their vision to life. With deep domain expertise and a future-proof philosophy, we embed innovation and agility into our clients'organizations—delivering measurable value and lasting change across industries, and around the world. Together, with over 26,000 employees in 25 countries, we build for boundless impact—touching billions of lives in the process. Visit us at:

ust.com