WHITE PAPER

Rethinking data platforms to enable digital transformation

latentview

Actionable Insights • Accurate Decisions

Evaluating the changing technology landscape

Organizations are enthusiastically transforming their operational processes and infrastructure with a razor-sharp focus on the customer at its core. The success of a business now truly depends on the rate and depth of analytics within its organization in a bid to serve customers a more intimate and targeted experience. The accrescent adoption of the digital technology landscape by users and businesses to buy and sell has convinced organizations to direct efforts on instituting appropriate data infrastructure and advanced analytics processes internally to cater to the ever-evolving digitally driven customer of today. At LatentView Analytics, we specialize in driving digital transformation for our clients by rethinking their data platforms and leading internal business processes to grow through their analytical process adoption curve.

A data science terrain, as we understand through our years of industry experience, is a potpourri of data, analytics, infrastructure, security and services directed towards the betterment of internal processes, ultimately aiming towards customer satisfaction. At the intersection of data, infrastructure and analytics lies modern data platforms, which, when built for specific business needs and securely presented for adoption, can yield returns manifold vis-à-vis a traditional data warehousing set-up.

Identifying the weakest link in your data warehouse

The traditional data warehouse set-up has accommodated conventional business use cases like reporting and data extraction. To power and support the diverse needs of different data user personas, the entire data architecture has to evolve and support multiple analytical workloads such as interactive, real-time and advanced analytics.

Cloud and other newer technology paradigms have led to a sharp growth in adoption of data to power everyday business process. While migrating to the cloud from on-premise data warehouses eventually will provide the estimated cost and efficiency benefits, it is not directly solving the core problems associated with engineering analytics platforms. In order to build and provision a platform that holds the central ground truth and enables an accountable way of delivering analysis, one must ensure business continuity by providing necessary support to data consumers.

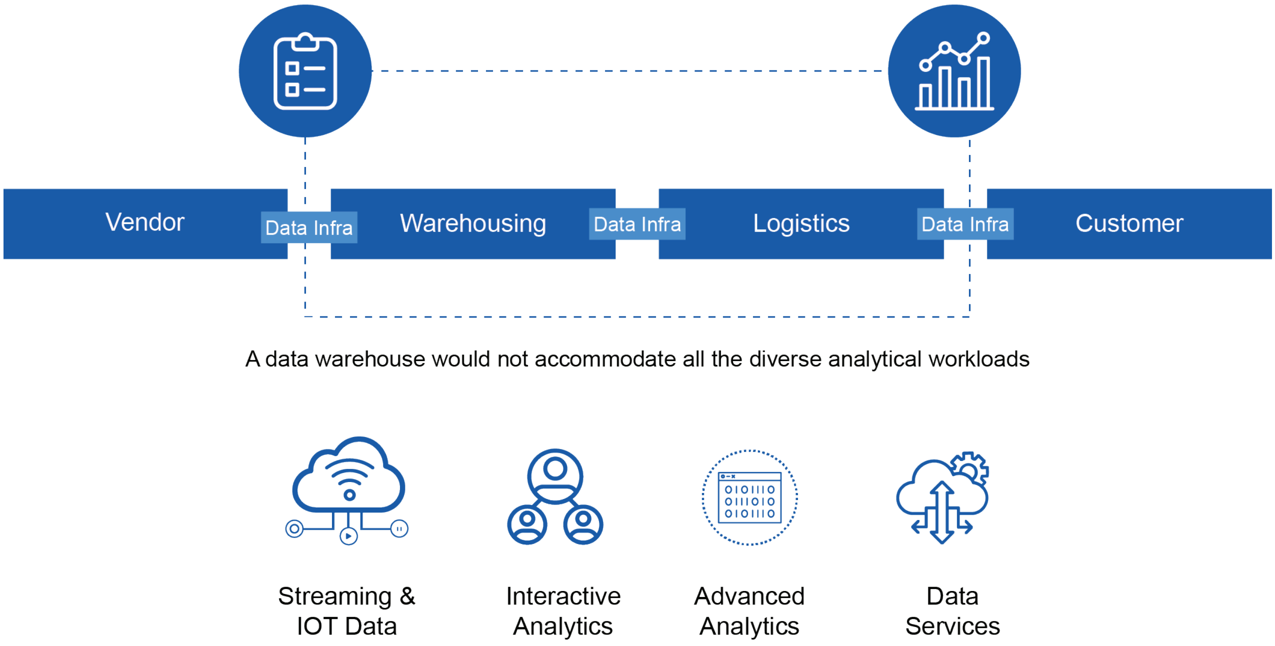

Figure 1: Diverse Analytical Workloads

In this regard, organizations can look at short-term solutions like fine tuning the warehouse to accommodate different/diverse workloads. However, this will eventually lead to the throttling of resources and restraining access. A typical concern of a data analyst would be to have sufficient horsepower and workspace to help narrate a convincing business point of view. This would warrant provisioning custom workspaces and additional capacity to support their requirements. Business users and managers would like to explore data visually and analyze various relationship between metrics.

As data warehouses are not naturally designed to support such requirements at scale, business users have greatly relied on third party tools to serve themselves data and propel their analysis. Data scientists might want to build sophisticated and best-of-breed analytical models for various business use cases. As they reach out to pre-process and extract years of transactional and dimensional data, the warehouse will naturally come to a grinding halt. Data scientists are then forced to down-sample data as much as possibly in a stratified way and continue with benchmarking their advanced analytics approaches to multi-dimensional and nonlinear business problems.

This creates an ecosystem where multiple versions of truth are maintained as data has to be pumped into multiple destinations for consumption by different business groups. A bigger challenge in such data architectures is a lack of processes to benchmark, qualify and productionize advanced analytics models.

In order to support the newer analytical workloads, the data ecosystem gets fractured and as a result, leads to no central ground truth, an inability to collaborate across different data ecosystems, constraints in productionizing advanced analytics solutions at scale and above all, a lack of accountability and user-friendly way to consume data.

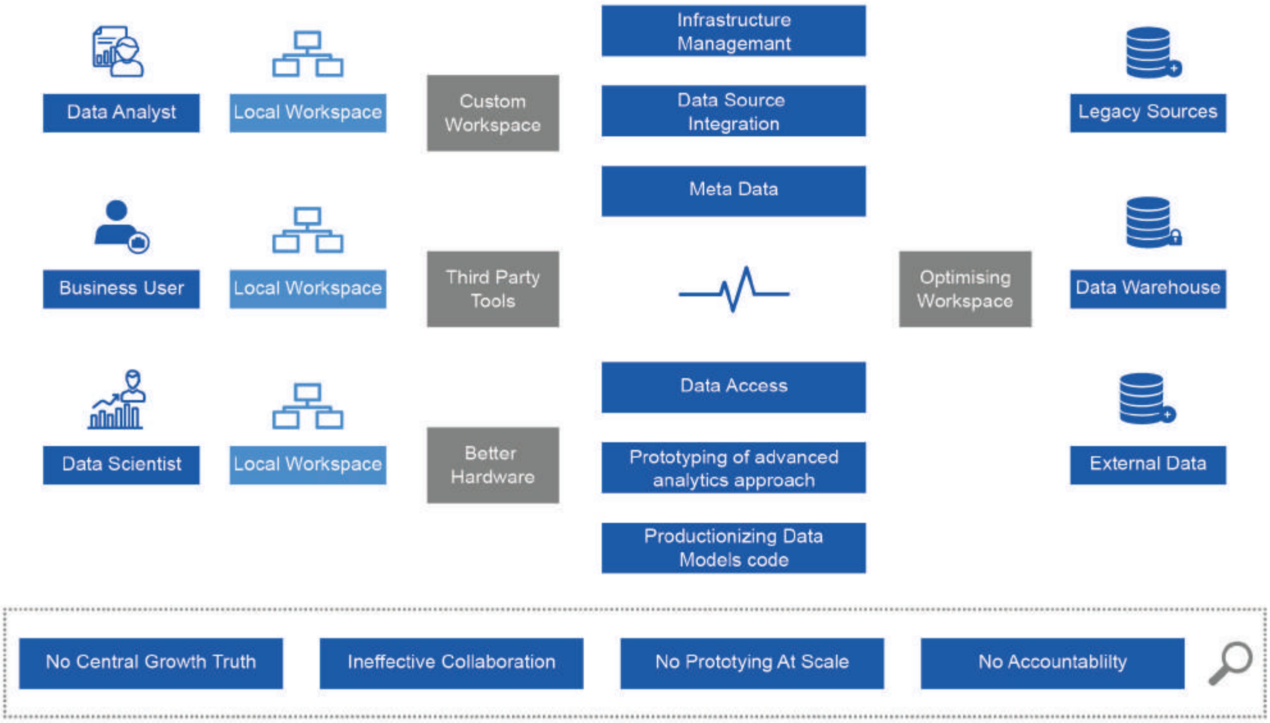

Figure 2: Current Enablement processes for data and analytics requirements

Why you need to rethink your data platforms

Any infrastructure that deals with analytics should be proactive and should be built to reduce lead time to insights. As the data moves through its nozzles, the systems and infrastructure should harness the value out of the same. The need to be at the right place at the right time with the right focus makes all the difference. But, fundamentally you have to be data hungry!

The world is indeed one big data problem. Industries have certainly moved from data insufficiency to data inundation. This poses a huge concern for organizations building and evolving their analytics infrastructure. To be able to mine from the data deluge to build a customer-centric business process is the most critical need and ask.

> ***“It is a capital mistake to theorize before one has data. Insensibly, one begins to twist the facts to suit theories, instead of theories to suit facts.”.***
>
> *Sherlock Holmes*

What would it take to build an ideal data ecosystem that supports multiple analytical workloads without compromising on the central ground truth? What will be the framework of a workload aware data ecosystem? What are the core problems that will be solved and what are the other challenges that will arise as we set out to implement a fundamental, but significant change within the organization that will look to embed analytics at the core of all its business processes? These imperative questions need to be answered before setting up the ideal data ecosystem.

Challenges of multiple data sources

Our objective is to architect a central ground truth that can accommodate any type of compute requirements to serve a business use case. In order to scale the infrastructure and user consumption, we should also infuse accountability and enforce governance into the overall setup. As data is gathered in the ecosystem from disparate sources, it brings with it, four primary challenges – Velocity, Variety, Volume and Veracity. Any analytics infrastructure should be built to handle these four problems. The extent to which those challenges are addressed qualitatively determines the success of the infrastructure. Different businesses focus on addressing different priorities. As organizations mature in their analytics and business processes, they have to invariably address all the four problems at some point. **That being said, at the core of all the problems is veracity.**

Building a single source of true data

In a bid to build analytics capabilities, different teams tend to curate data from diverse sources, bake in their domain specific pre-processing, filters and post processing, apply rules, analyze and even end up building machine learning models to answer their business priorities. In this process, invariably, different business teams tend to build their version of a “truthful” source and continue to build and benchmark results on their version of “truthful” data. This leads to data fragmentation across different business domains. This fragmentation problem is invisible until a need comes in to unify business processes, collaborate and build data solutions that cuts across different business domains.

In a bid to build analytics capabilities, different teams tend to curate data from diverse sources, bake in their domain specific pre-processing, filters and post processing, apply rules, analyze and even end up building machine learning models to answer their business priorities. In this process, invariably, different business teams tend to build their version of a “truthful” source and continue to build and benchmark results on their version of “truthful” data. This leads to data fragmentation across different business domains. This fragmentation problem is invisible until a need comes in to unify business processes, collaborate and build data solutions that cuts across different business domains.

Origins of a central data repository

Let’s attempt to define the traits of a central data repository.

Workload/Compute agnostic for	• Metrics Computation, Reporting • Ad-hoc Analysis • Advanced Analytics • Real-time Systems
Secure access via	• Client specific integrated authentication/authorization
Ground truth dataset available as	• Latest version of data • Cumulative history (Immutable)
Optimized for compute and storage as	• Columnar • Compressed • Partitioned
Discoverable via	• Tagged metadata
Systematic refresh via	• Configurable inputs
Advanced Analytics Workbenches for	• At-Scale prototyping • Continuous integration and deployment of machine learning models • A/B testing and Integration of AI/ML results into application
Intuitive summary and ease of use	• Statistical summary and Anomaly identification • Intuitive interactive visualizations

How to build an ideal data ecosystem

At LatentView Analytics, we work with clients across banking, finance, CPG, energy, retail, healthcare, media, telecom, technology and partner with global academic institutions to build solutions that are state of the art and relevant to business. In order to power the next wave of analytics-driven decision making at scale, we do see a strong need for an organized and well-managed data repository for many of our clients. By considering the ideal characteristics of a central data repository discussed above and assessing the analytics maturity of our prospects and clients, we have built a klotski styled ideal data engineering stack that will deliver, manage and run the modern analytical engine to fuel the growth of organizations.

Our approach to engineering an ideal stack depends on the level of analytical maturity within our client organizations. The Analytical Maturity Assessment is custom designed for different business verticals and is very useful to propose the right approach, appropriate technical components and outcome bound priorities based on short and long term strategic goals.

A unified data fabric will be the central data repository that will be periodically refreshed from various internal and external sources. The refresh by design is configuration driven in order to maximize the re-use of code and environment setup. The key aspect to note here is the proactive isolation of production and development environments. This is often overlooked in many data engineering setups and eventually leads to ineffective analysis and incomplete benchmarking. As data ecosystems evolve, richer functionalities are required to support easier adoption. Exposing business and technical metadata makes data discoverable over a search and additional information such as lineage and definitions help build trust between the data infrastructure and data consumers.

In a contemporary set-up, there is a great need for self-serve analytics which is used operationally by thousands of users within an organization for every day decision making. It is very important in such scenarios for those data hungry users to feed descriptive statistics and intimation about anomalies to help them form a better intuition about the data. It will save hours of analysis by hinting about potential data differences and metric deviations expected in their reports and summary. In addition to statistics about data, it is also very critical to expose infrastructural and operational metrics to analyze and fine tune the performance of the entire ingestion and analysis process.

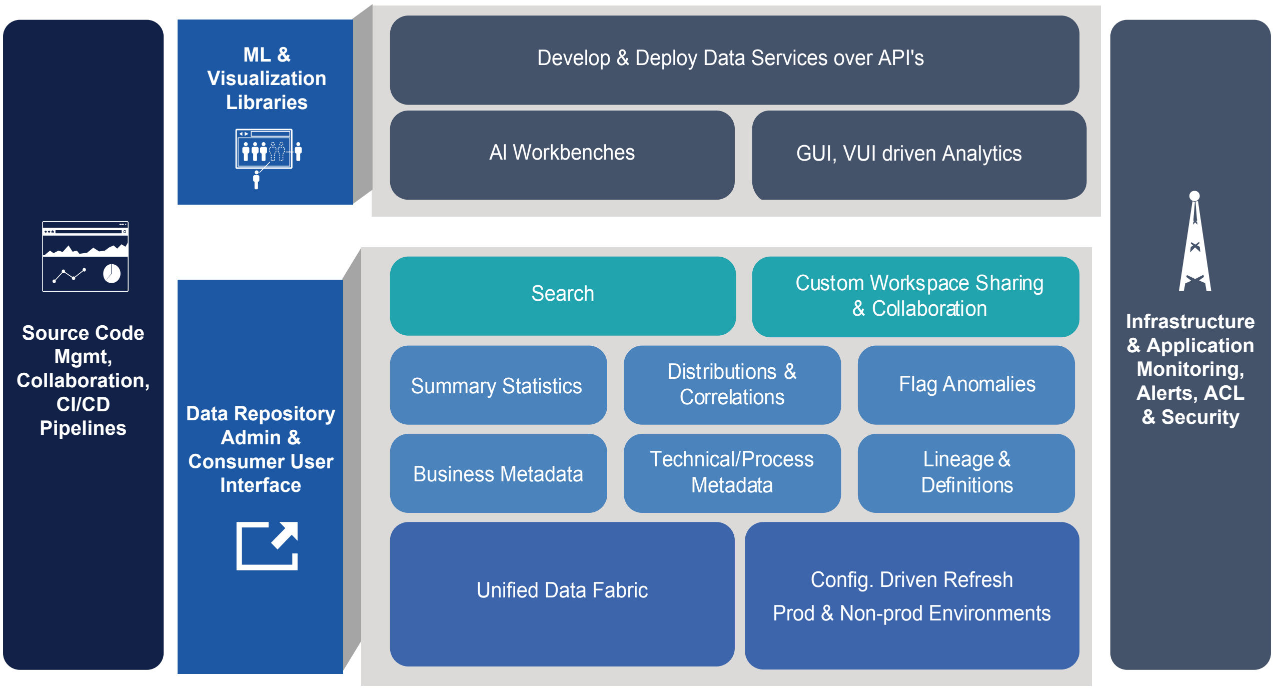
Source Code Mgmt, Collaboration, CI/CD Pipelines
ML & Visualization Libraries
Develop & Deploy Data Services over API's
AI Workbenches
GUI, VUI driven Analytics
Data Repository Admin & Consumer User Interface
Search
Custom Workspace Sharing & Collaboration
Summary Statistics
Distributions & Correlations
Flag Anomalies
Business Metadata
Technical/Process Metadata
Lineage & Definitions
Unified Data Fabric
Config. Driven Refresh
Prod & Non-prod Environments
Infrastructure & Application Monitoring, Alerts, ACL & Security

Figure 3: Ideal Data Engineering Stack

There are three major data-powered actions that are commanded out of a new age analytics system. They are as follows:

- Advanced Analytics Workbenches for users to prototype approaches at-scale
- Interactive analytics (GUI based, Voice powered and Conversational)
- Data as a service (DaaS) in a secure and scalable fashion

These are precisely the requirements that break the traditional monolithic warehouse. In order to accommodate the same, we have to invest in setting-up a continuous code integration, testing and deployment platform that makes the entire process accountable and manageable at-scale. With an all new set-up and process, the harder problem to handle is rollout and adoption without impacting business continuity.

This can be achieved by very closely aligning with the business priorities and involving all the business representatives as part of the design and implementation. There could be steep learning curves involved with respect to user onboarding and technology adoption. This concern can be mitigated by conducting deep-dive and hands-on workshops. Last but not the least, incentivize users by public recognition of adoption, document and promote necessary content and engage by clearly explaining the overall business benefit that the organization derives by moving to a cleaner and scalable engineering analytics stack.

Evaluating Technical Components for Data Engineering & Advanced Analytics Stack

What is the point of engineering infrastructure if it is not going to be put to good use? Any well thought out and built engineering stack should serve its purpose by helping analysts and business users build and deliver value to customers in the most effective and efficient way. In order to reap benefits of such a platform, we have earlier identified various logical components that forms its core. We will take a closer look into each of those logical components and identify technology choices available for each of those. Let's look into each of the components and various technology choices in detail.

> ***"If we have data, let's look at data. If all we have are opinions, let's go with mine."***
>
> *Jim Barksdale*

Designing the unified data fabric

This is both a logical and physical data layer that can accommodate a wide spectrum of data storage solutions ranging from DBs to File Systems to Object Stores. The challenge is to ensure that data is consistent across all the different technologies used. We prefer to consolidate the data fabric to as minimal a set-up as possible and diversify later in the compute needs. This will avoid creating multiple data silos for different requirements.

Components	Technology Choices
Data Warehouses - MPP	Redshift, MemSQL, CosmosDB, SQL Server
File System	EFS, NFS, HDFS
Object Storage	S3, Blob Storage, Azure Data Lake Store

Configuration driven refresh and environments synchronization

A critical aspect of a data ecosystem is the data ingestion strategy. In addition to the traditional tools available, we strongly propose consolidating the code and infrastructure set up to as much as possibly leverage configuration and generalized code to ingest data into the unified data fabric. In our case, we have custom built ingestion modules to read configuration and dynamically ingest data from respective sources. Ingestion can be categorized into batch, mini batch and continuous.

For each of those, there are different technology choices available

Components	Technology Choices
Workflow orchestration tools and ETL tools	Pentaho, Informatica, Talend, Oozie, Airflow, Azkaban, Apache Nifi
Configuration Store	ElasticSearch, NoSQL
ETL Platform(Batch & Mini Batch)	EC2, ECS, Containers, Hive, Spark (EMR/ DataBricks/Qubole), AWS Glue
ETL/Data Platform(Continuous)	Kinesis, Kafka, AWS Lambda, Azure Functions

In addition to refreshing data into the fabric and making it available, it is equally important to isolate production and development environments to enforce security and sanity to prevail over the data repository.

The data successfully refreshed in production has to be synchronized into a development environment with similar set-up. The analysts and business users can access the data in the development environment to prototype and test reports, perform analytics and build models. The data is also QA'ed on every sync to review any mismatch in data if any.

Metadata and search

Metadata is an important differentiator when it comes to our data ecosystem stack. It is integral to the way datasets are onboarded on to the data fabric. It is critical and mandatory to collect and persist business and technical metadata about the datasets ingested in to the ecosystem. The consumers of data would be able to search, discover and filter datasets relevant to their problem scenario by greatly leveraging the metadata tagged to respective datasets.

In addition to business descriptions, column and data type information, we also capture the lineage of each metric and table data to facilitate easier debugging and understanding of computations beneath the dimension or fact. The inter-linked nature of metrics and its respective computations can be explored via graph databases.

Components	Technology Choices
Metadata Extraction	Custom SQL Parser, Third Party Tools
Lineage Extraction	Third Party Tools (Collibra, Alation)
Business Metadata	Forms, Custom UI & CSVs
Data Catalog As a Service	AWS Glue
Search	Elastic Search, SOLR
Graph	Neo4j

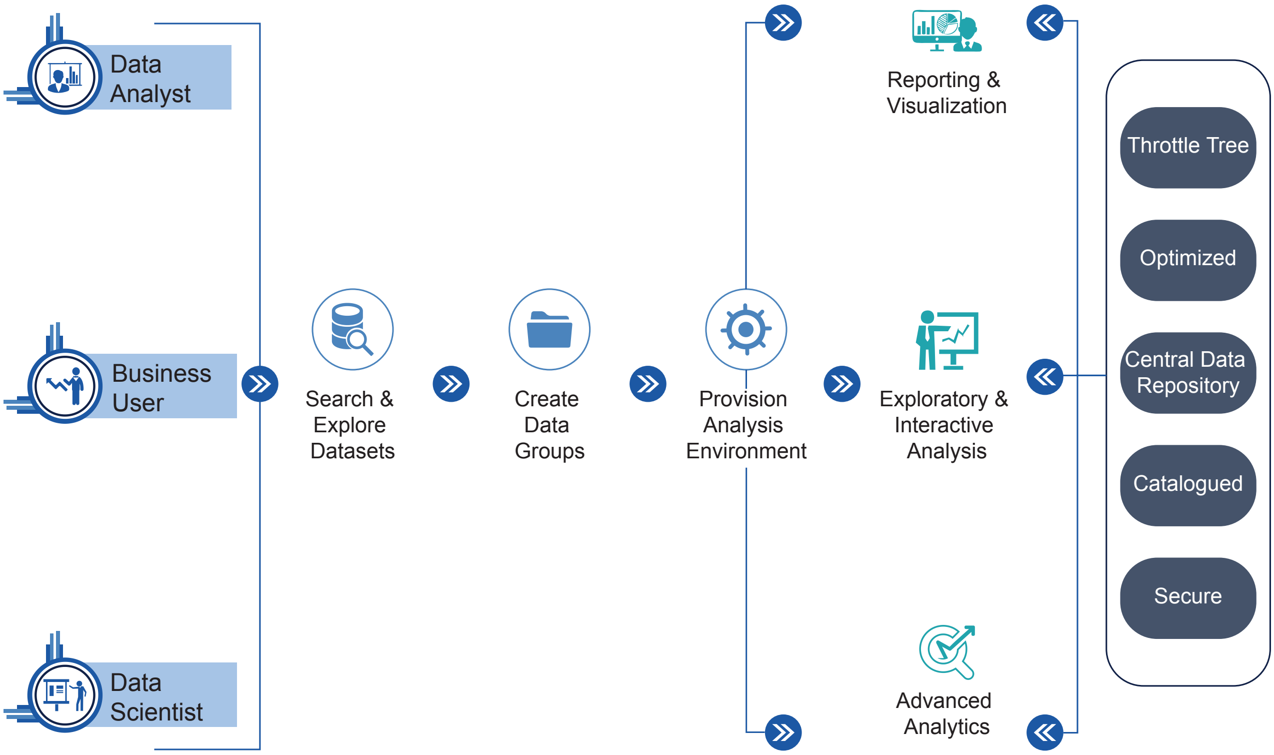

Figure 4: Data Consumption Process

Out of the Box: Data Statistics

Every analyst and consumer of data would prefer to have a first-hand summary of data using descriptive and exploratory analysis. Most of the questions sought after in the preliminary phase of data understanding is pretty consistent across different use cases. We have built a code-free data exploration engine for the analysts and business consumers of data. This platform leverages metadata and the underlying data fabric to pre-compute and store details about the distribution of data, anomalies, distinct values, correlation between key columns. This detail comes in handy for the analyst to form a better intuition about the data for respective analysis

Components	Technology Choices
Exploratory Analysis & Summary	Amplifyr (Custom built tool by LatentView)

Code, custom workspace and collaboration

In order for any data ecosystem to be adopted by all team members, it is important to have support while developing and maintaining code, reviews, commits and custom workspaces where users can run, evaluate and optimize their code. There are a whole bunch of code management solutions integrated with documentation and project management utilities available. The subtle challenge then is to provide an appropriate workspace or prototyping platform for users to test their code for further action.

While simple initial development can happen in respective local machines of users, it is absolutely necessary for them to prototype their solutions at scale before they can be approved and deployed.

In addition to code management, we need to enforce rigorous code review and sign off process which when approved would automatically be built, tested, integrated and deployed to appropriate destination paths. We leverage CI/CD pipelines to handle the automated deployment process.

Components	Technology Choices
Code & Project Management	Github, JIRA, bitbucket, gitlab, S3i
Documentation	Confluence, gists using Markdowns
Custom Workspaces	S3, Dedicated folders in shared file systems
CI/CD	Jenkins, Cloud Formation

Access control, data security and governance

Monitoring and alerting on usage and access ensures a safe and reliable ecosystem. This phase of enforcing access control, security and governance involves stakeholders across the organizations from IT, business, data engineering, network and security.

Components	Technology Choices
Access Control	Apache Ranger, ACLs, IAM
Data Security	Custom Third Party tools as per org priorty
Monitoring & Alerting	Grafana, StatsD, Nagios, CloudWatch, Twilio, SNS, SES, Logstash, Kibana, Elastic Search
Audit	Resource and Dataset level tagging

Advanced analytics workbenches

We also allow for users to provision a custom analysis environment subject to the business needs. We have understood from our customers that there is a need for notebooks for data science code development as they help in quickly iterate on code and debug faster.

These notebooks can be configured to sync the code developed with the code management utilities listed in the previous section.

Components	Technology Choices
Notebooks	Jupyter on EC2, Zeppelin on EMR/Spark, RStudio on EMR/Spark

Data services and serverless stack

Nowadays, teams within organizations have their own custom applications that consumes data from the central data repository and automatically perform intelligent operations on the same. In order for different applications to consume data programmatically, we have to expose the data over an API. The load and scale is not very predictable as different applications can consume different data volumes and types based on its respective requirements. This drives most of the organizations to evaluate and deploy a serverless stack using cloud based solutions available. It is also very important to evaluate an appropriate backend to support the scale and concurrency demand of an application before identifying the technical components.

Components	Technology Choices
Backend DB	MySQL, MemSQL, Redshift, PostgreSQL
Cache	MemCache, ElastiCache, Redis
API	API Gateway, Custom NodeJS apps
Processing	AWS Lambda, Custom NodeJS apps, Python
Logging & Monitoring	MySQL, DynamoDB, Cloudwatch, ELK stack
Queue	SQS, Celery, RabbitMQ, Redis

Data management solutions

There are few data management solutions available. These solutions accommodate some of the important components discussed above. We evaluate these tools and try as much as possibly to customize the existing open source code before developing anything grounds up. These solutions listed below are available for evaluation and is better to try it out, benchmark before making a choice.

- Kylo
- AWS Data Lake
- Azure Data Lake

Engineering Data ecosystems the right way strategically aids organizations to reduce their lead time to insights. The past decade has seen tremendous growth in the engineering landscape and the promise of providing better customer experience and optimized business process is brighter than never before. The challenge resides in careful understanding of business requirements and chiseling the engineering stack to serve the data consumption pattern.

How LatentView Analytics can help accelerate your Data Engineering capabilities

At LatentView, we have built and delivered centralized data repositories using the above solutions and are currently deployed at scale, ingesting and analyzing terabytes of data in our client environments enabling them with actionable insights.

To know more, reach out to sales@latentview.com

About the author

Karthick Hariharan
Engineering Manager

Karthick Hariharan, Engineering Manager, LatentView Analytics, is a hands-on leader in building and managing data platforms for analytics at scale. He plays a key role in mentoring teams and leading them to build sophisticated pipelines for data processing and analysis. Leveraging the power of advanced analytics, he has built pipelines for processing unstructured to structured data, financial analytics, product metadata analysis and near real-time search indexing. With over 9.5 years of industry experience, he has rich exposure to systems, networking, cloud, data platforms, E-commerce, search and recommendation systems.

About: LatentView Analytics

LatentView Analytics is a leading global data and analytics service provider helping companies turn data into actionable insights to gain competitive advantage. As a trusted analytics partner to the world's most recognized brands, LatentView solutions provide a 360-degree view of the digital consumer, fuel machine learning capabilities and support artificial intelligence initiatives. LatentView’s success is driven by a commitment to deliver unrivalled analytics solutions that enable Fortune 500 companies in the retail, CPG, BFSI, high tech, healthcare and other sectors to predict new revenue streams, anticipate product trends, improve customer retention, optimize investment decisions and turn unstructured data into a valuable business asset. LatentView has offices in Princeton, N.J., San Jose, Calif., London, Singapore and Chennai, India with more than 600 employees globally.

For more information, please visit www.latentview.com or follow us on LinkedIn.