Unit - 04

Scalability and Changes to the legacy processes

One process that needs to be changed is the process of configuring and maintaining workspace for analytic professionals.

Traditionally, this workspace was on a separate server dedicated to analytical processing.

In-database processing is becoming the new standard.

In order to take advantage of the scalable in-database approach, it is necessary for analysts to have a workspace or sandbox residing directly within the database system.

In the big-data world, a MapReduce env. will often be an addition to the traditional sandbox

Analytic Sandbox

databases → building and deployment of advanced analytic process.

In order for analytic professionals to utilize an enterprise data warehouse or data mart more effe they need the correct permissions and access to so. An analytic sandbox is the mechanism doing so. If used app. an analytic sandbox can be one of the primary drivers of value in world of big-data.

"Sandbox" → sandboxes that children play in. Within a sandbox, children can create anything they like. They can reshape the sand as per their desires.

Similarly sandbox in the analytics context is a set of resources that enable analytic professionals to experiment and reshape data in whatever fashion they need to. (Agile Analytics, data lab).

→ In-depth analysis → answer critical business quest.
→ Ideal for data exploration.
→ development of analytical processes / prod^{n} processes
→ proof of concepts and prototyping.

Once things progress into ongoing, user managed process or production process, then sandbox should not be involved.

No permanent data – build the data needed for the project.

When project done – delete the data.

Benefits of analytic Sandbox

(1) Independence: professionals will be able to work independently w/o continually going back and asking for permissions.

(2) Flexibility: Analytic prof. can use whatever BI, Statistical Analysis, visualization tools that they need to use.

Programming Model.

Advantages & Disadvantages.

Disadv.

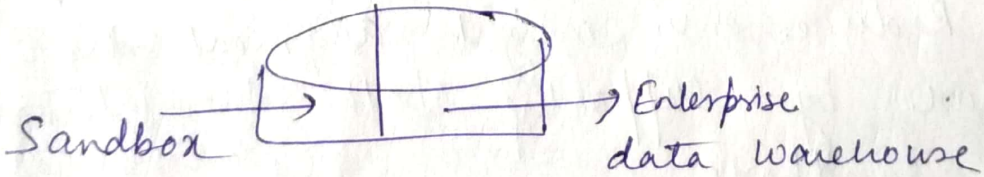
small data.

Efficiency: leverage existing enterprise data warehouse or data mart, w/o having to migrate or move data

(4) Freedom: Can reduce focus on the administration of systems and babysitting of prodⁿ processes by shifting those maintenance tasks of IT.

(5) Speed: Due to ~~massive speed~~ parallel processing. Enables rapid iteration and ability to 'fail fast' and take more risks to innovate.

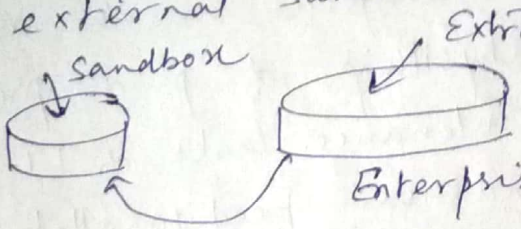
Internal Sandbox: A portion of an enterprise data warehouse or data mart is carved out to serve as the analytic sandbox. Sandbox is physically located on the production system. Sandbox database is not a part of production database.



Big Data → Map-Reduce Environment is added.
The Map Reduce Env. will require access to the internal sandbox.
Leverages existing I/O resources and infrastructure ready in place.
Join production data with sandbox.

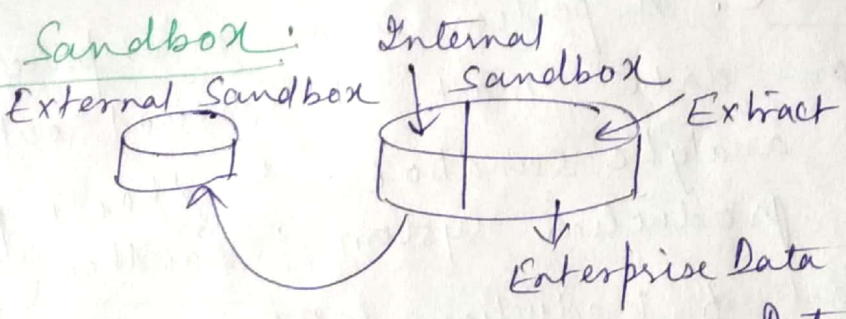
Simple Program... Advantages ... fit for small data...

The External Sandbox : A physically separate analytic sandbox is created for testing & dev. of analytic processes. Relatively rare to use an external sandbox.



It will have no impact on other processes. which allows for flexibility in design and usage.

Hybrid Sandbox :



A single Map Reduce env. might augment the hybrid sandbox by supporting both the internal and external sandboxes.

Analytic Data Sets : ADS is the data that is pulled together in order to create an analysis or model. It is the data in the format required for the specific analysis at hand.

ADS is generated by transforming, aggregating & combining data. There will be one record per customer, location, product or whatever type of entity is being analysed. Bridge the gap. b/w efficient storage and ease of use.

lytic Point Solutions : Analytic point soln are software packages that address a very specific, narrow set of problems. Typically they focus on a set of related business issues, and they often sit on top of analytical tool suites. Examples of point soln include price optimization applications, fraud applications and demand forecasting app.

Point solutions are built and configured to constrain a user to action that are appropriate. Once the solutions are configured and set up by experts, they enable automation of many tasks so that a power user is able to monitor the tool's output and make sure everything is working okay.

Enterprise Analytic Data Sets :

EADS is a shared and reusable set of centralized, standardized analytic data sets for use in analytics.

It condense 100s or 1000s of variables into a handful of tables and views which will be available to all analytic professionals, applications & users.

Various analytic process can share the same, consistent set of metrics.

Evolution of Analytic Tools and Methods

Ensemble methods | Commodity modeling | Text Analysis

Ensemble Method: Instead of building a single model with simple technique, multiple models are built using multiple techniques. Once results from all of the models are known, all the results are combined together to come up with a final answer. The process of combining the various results can be anything from a simple average of each model's predictions to a much more complex formula.

The power of ensemble methods lie in comparing various technique to get a single answer.

Commodity Models: The model that is produced rapidly and with less concern for squeezing out every ounce of predictive power. Done with stepwise analysis procedure

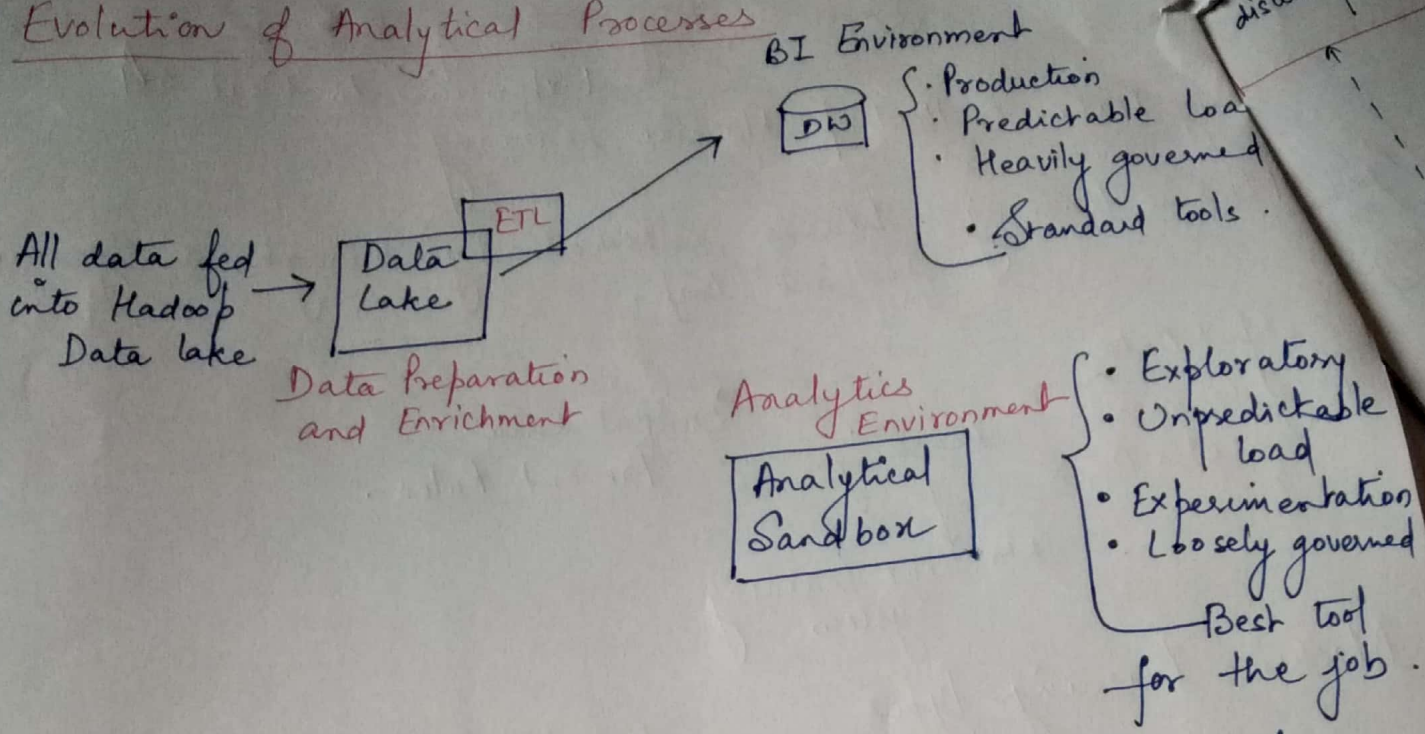
The goal is not to get the best model but to quickly get a model that will lead to a better result than if there had been no model at all.

Enable the application of advanced analytic to a much wider scale within an org.

Text Analysis: Analysis of text and other unstructured data. It takes some sort of text as input. This text can be written material like an e-mail, transcribed material. A lot of big data falls in this category. Takes text as input. Organizations are starting to have more text and unstructured data available to them, than they do traditional, structured data. Text is a very common type of big-data and text analysis tools and methods have come a long way. Popular commercial text analysis tools include those offered by Attensity, Clarabridge, SAS and SPSS

Once the text is parsed into its components, there are methods that will help identify the sentiment or meaning of those components and find trends within them

Evolution of Analytical Processes

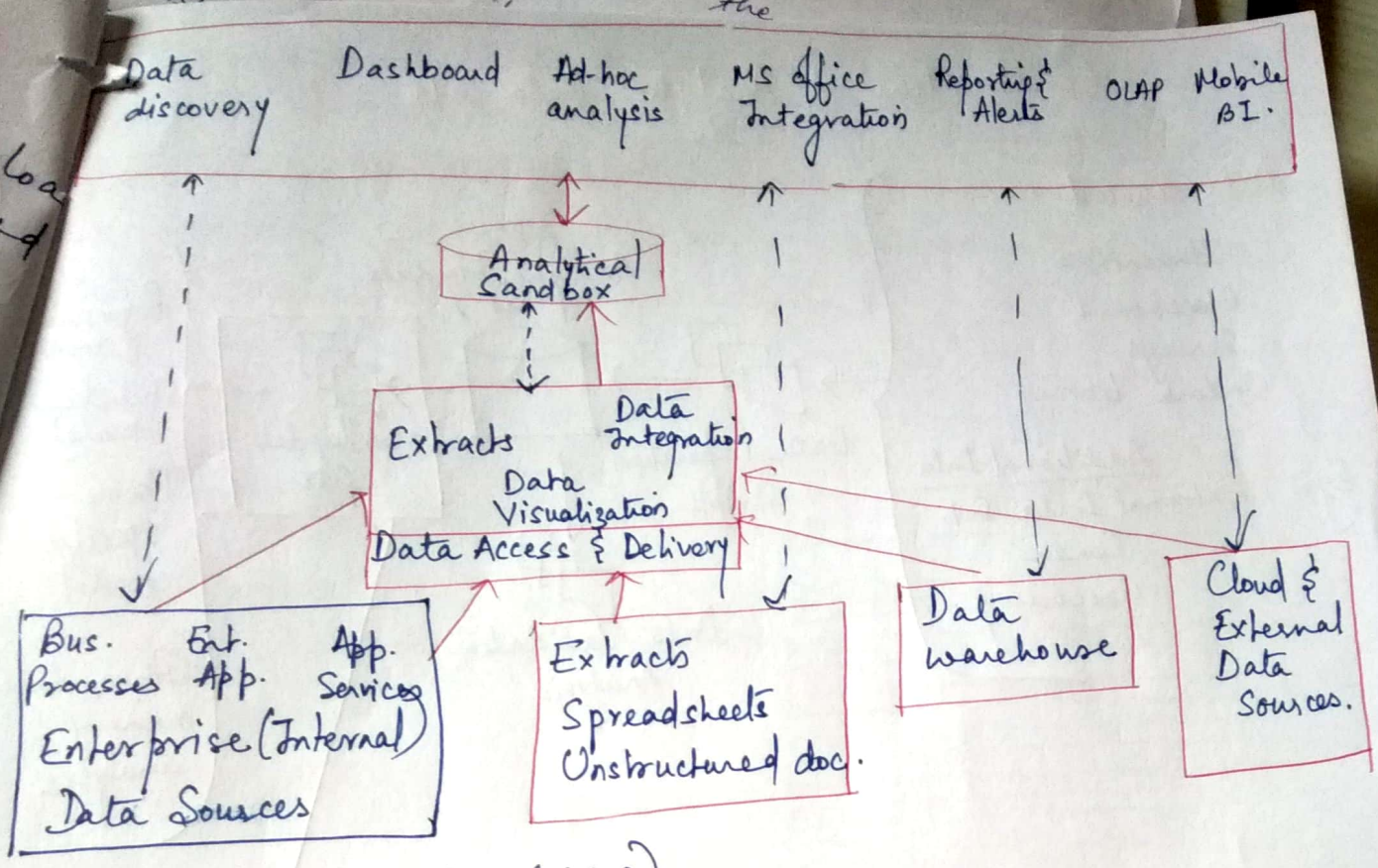


Upgrading technologies won't provide a lot of value, if the same old analytical processes remain place.

(a) Change the process of configuring & maintaining workspace. The Analytic Sandbox

(b) Consistently leverage a database platforms through a sandbox Enterprise Analytic Data Set (EADS)

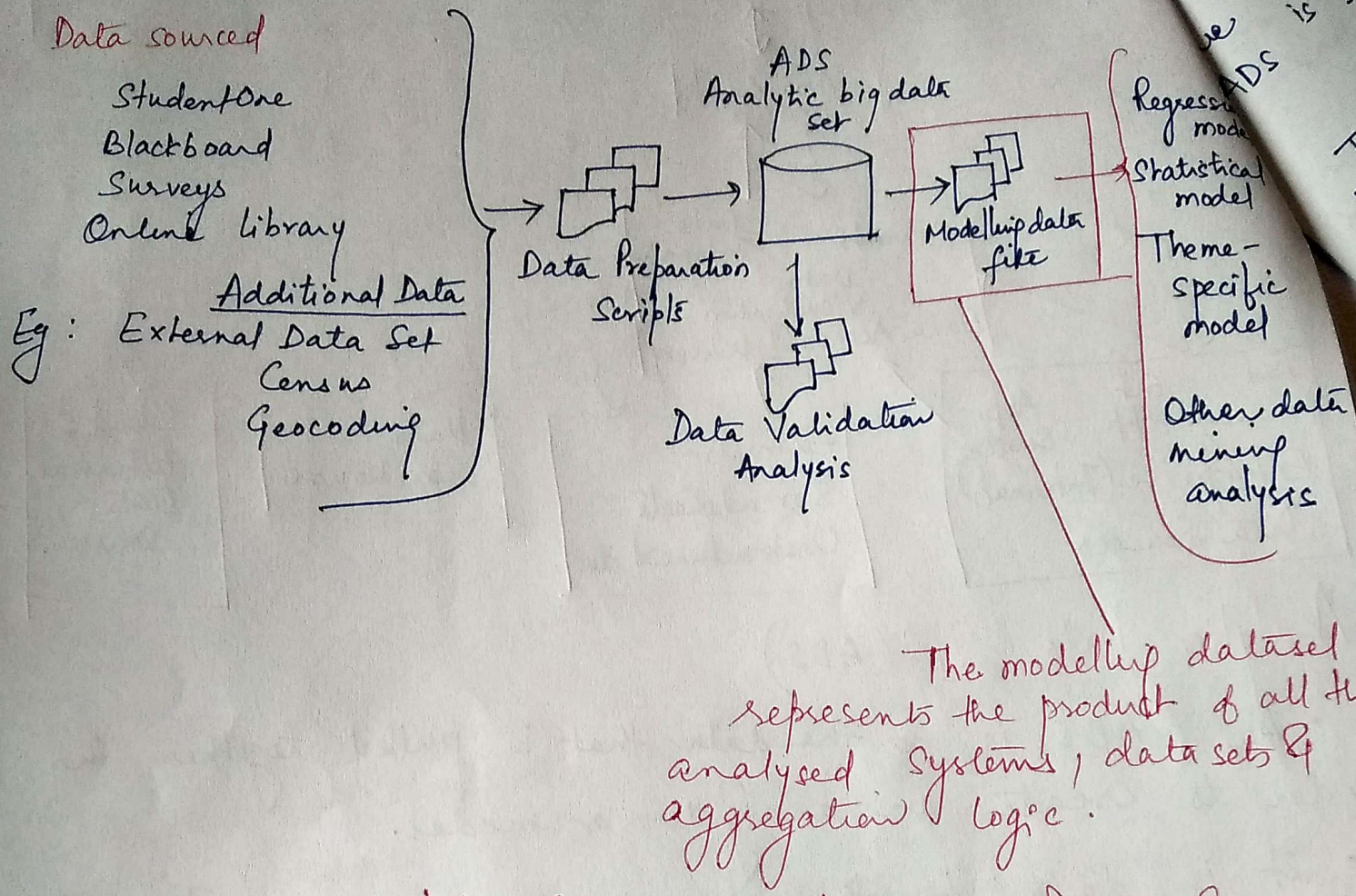
(c) Necessary to keep scores up to date on a daily Embedded Scoring.



Analytic Data Set (ADS)

- An ADS is all the data that is pulled together in order to create an analysis or model.
- It is data in the format required for the specific analysis at hand. An ADS is generated by transforming, aggregating and combining data.
- It is going to mimic a denormalized, or flat file structure. What this means is that there will be one record per customer, location, product, or whatever type of entity is being analysed. The analytic data set helps to bridge the gap between storage and ease of use.

Construction of Analytic Data Sets



The modelling dataset represents the product of all the analysed systems, data sets & aggregation logic.

DEVELOPMENT Vs PRODUCTION ANALYTIC DATA SETS

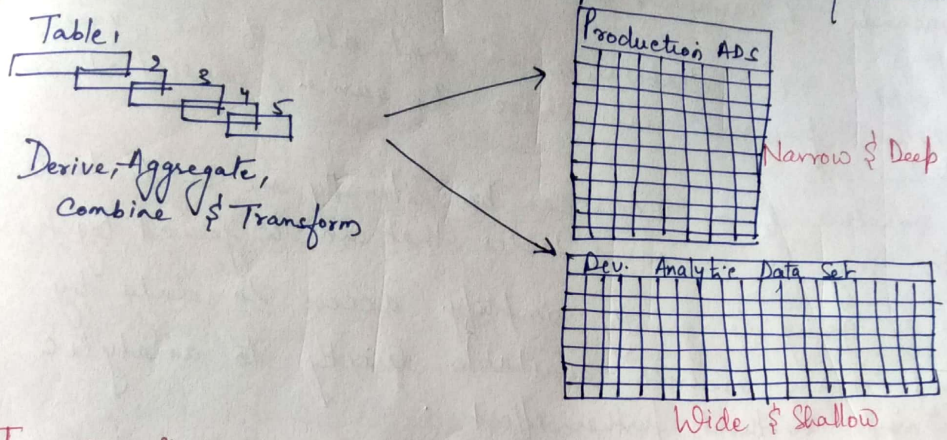
A production analytic data set, is what is needed for scoring and deployment.

A big difference here is that the scores need to be applied to every entity, not just a sample.

Every customer, every location, every product will need be scored. Therefore a production ADS is not going t very wide, but it will be very deep.

Example: When developing a customer model, an analyti professional might explore 500 candidates metrics for Sample of 1,00,000 customers.

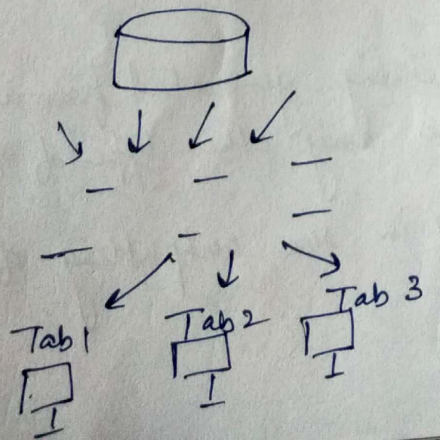
ADS is therefore wide and shallow.
when it comes to apply scores to customers in production, perhaps only 12 metrics are needed but they are needed for all 30,000,000 customers. The production ADS is therefore narrow but deep.



Traditional ADS: All analytic data sets are created outside of the database

Each analytic professional creates his or her own analytic data sets independently. Possibly hundreds of people generating their own independent views of corporate data.

Traditional ADS Process:
A dedicated ADS is generated outside the database for every project



Enterprise ADS

An EADS is a shared and reusable set of centralized standardized analytic data sets for use in analytics. What an EADS does it to condense hundreds or thousands of variables into a hand ful of tables and views.

An EADS is collaborative in that all of the various analytic processes can share the same, consistent set of metrics.

The structure of an EADS can be literally one wide table or it may be a no. of tables that can be joined together.

EADS is going to greatly simplify access to data by making many metrics available directly to analytic professionals without further effort.

Embedded Scoring

→ Embedded Scoring involves enabling scoring ~~Scoring~~ routines to run in the database so that users can leverage the models built in an effective scalable fashion.

→ Successfully implementing embedded scoring will include not just deploying each individual scoring routine, but also a process to manage and track the various scoring routines that are deployed.

Note that a 'score' can be something generated from a predictive model, or it can be any other type of output from an analytic process.

→ Analytic processes often result in the outputting of a new piece of information.

There are 4 primary components required to effectively manage all of the analytic processes an enterprise develops

There are commercially available tools to help with the model and score management or a custom solution can be built to address an organisation's specific needs

Analytic Evolution

Ensemble: Ensemble approaches are fairly straightforward conceptually. Instead of building a single model with a single technique, multiple models are built using multiple techniques. Once the results from all of the models are known, all of the results are combined together to come up with a final answer.

The process of combining the various results can be anything from a simple average of each model's predictions to a much more complex formula.

Certain types of customers, for example, may be scored poorly by one technique but very well by another. By combining intelligence from multiple models, a scoring algorithm becomes better in aggregate, if not literally for every individual customer, product, or store location scored.

Commodity: The model is produced rapidly and with less concern for squeezing out every ounce of lift or predictive power.

The goal is not to get the best model, but to quickly get a model that will lead to a better result than if there had been no model at all.

Text: One of the most rapidly growing methods utilized by organizations today is the analysis of text and other unstructured data source. A lot of big data falls into these classifications. Text analysis, as the name implies, take some sort of text as input. The text can be written material like an e-mail transcribed material such as a medical dictation, or even text has been scanned from a hard copy & converted to electric form like old courthouse records.

Point Solutions: A trend that has accelerated in the past decade is the availability of analytic point solutions. Analytic point solutions are software packages that address a very specific, narrow set of problems. Typically they focus on a set of related business issues, and they often sit on top of analytical tool suites.

Examples: price optimization application, fraud applications, etc.