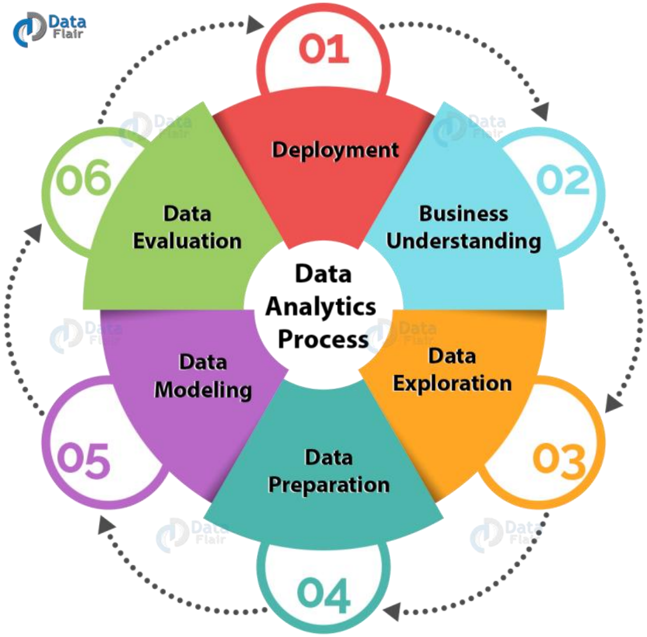
JECRC™
UNIVERSITY
BUILD YOUR WORLD

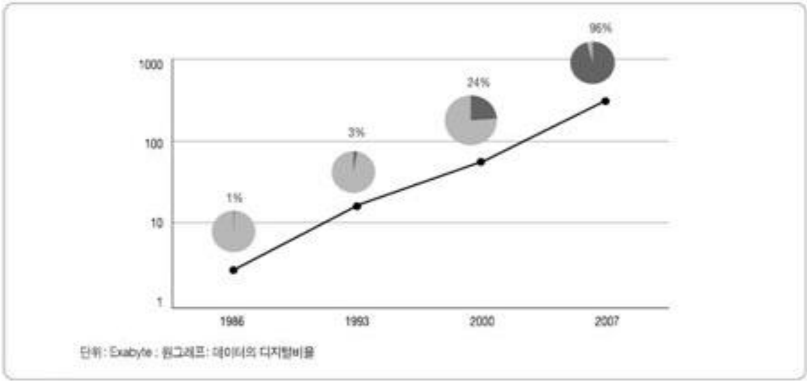


EVOLUTION OF ANALYTICAL PROCESSES

UNIT 4| BIG DATA ANALYTICS

Why Tame (Domesticating) the Big Data?

- The amount of data organizations process continues to increase



단위 : Exabyte ; 원그래프 : 데이터의 디지털비율

*출처: Hilbert & Lopez(2011) 재구성

[그림 1] 전 세계 정보량의 변화(로그 스케일)

The old methods for handling data won't work anymore

- Important technologies to tame the big data tidal wave possible

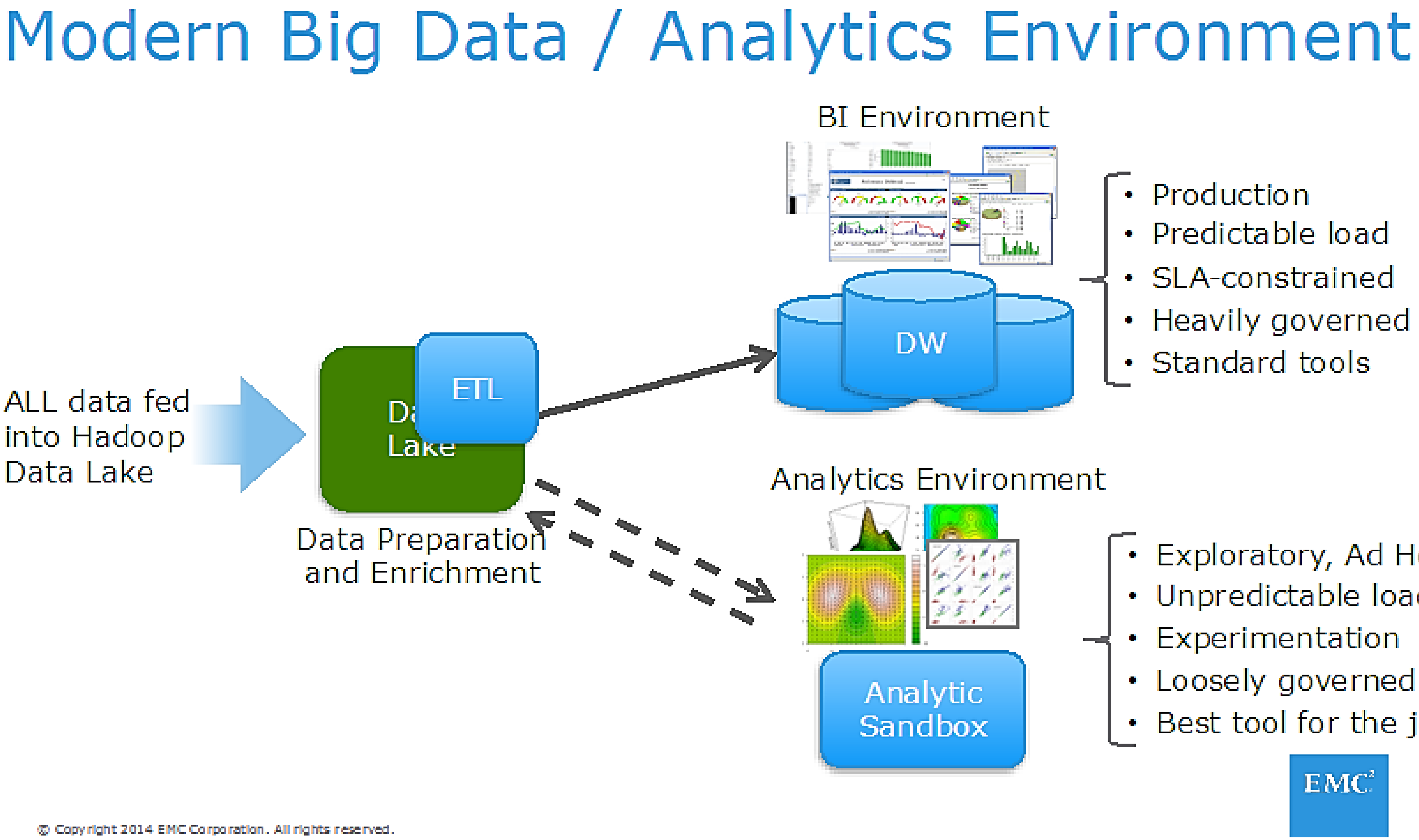
MPP

The cloud

Grid computing

MapReduce

Modern Big Data Environment



Modern Big Data / Analytics Environment

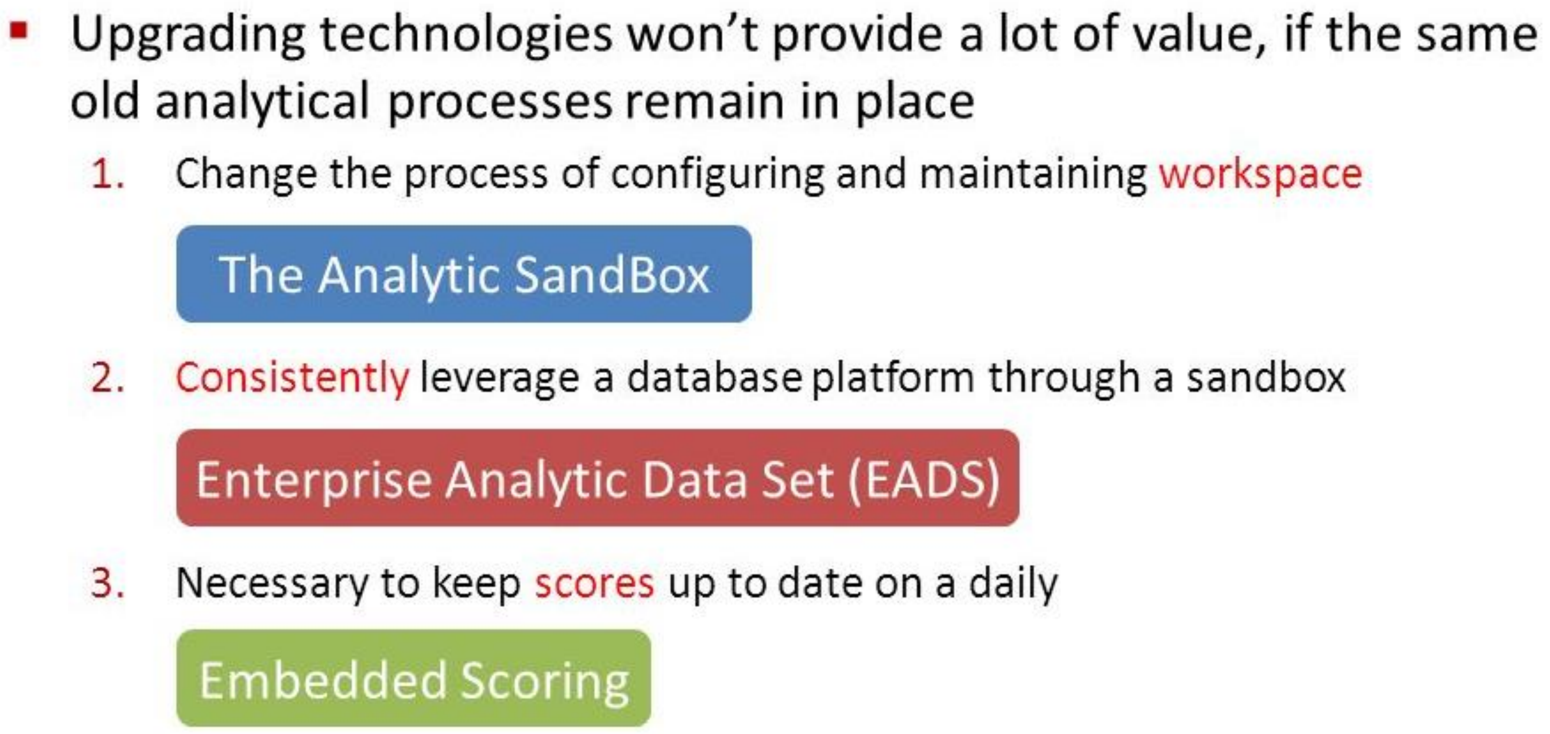
BI Environment

- Production
- Predictable load
- SLA-constrained
- Heavily governed
- Standard tools

Analytics Environment

- Exploratory, Ad Hoc
- Unpredictable load
- Experimentation
- Loosely governed
- Best tool for the job

Evolution Insights



- Upgrading technologies won’t provide a lot of value, if the same old analytical processes remain in place
  1. Change the process of configuring and maintaining workspace

     The Analytic SandBox

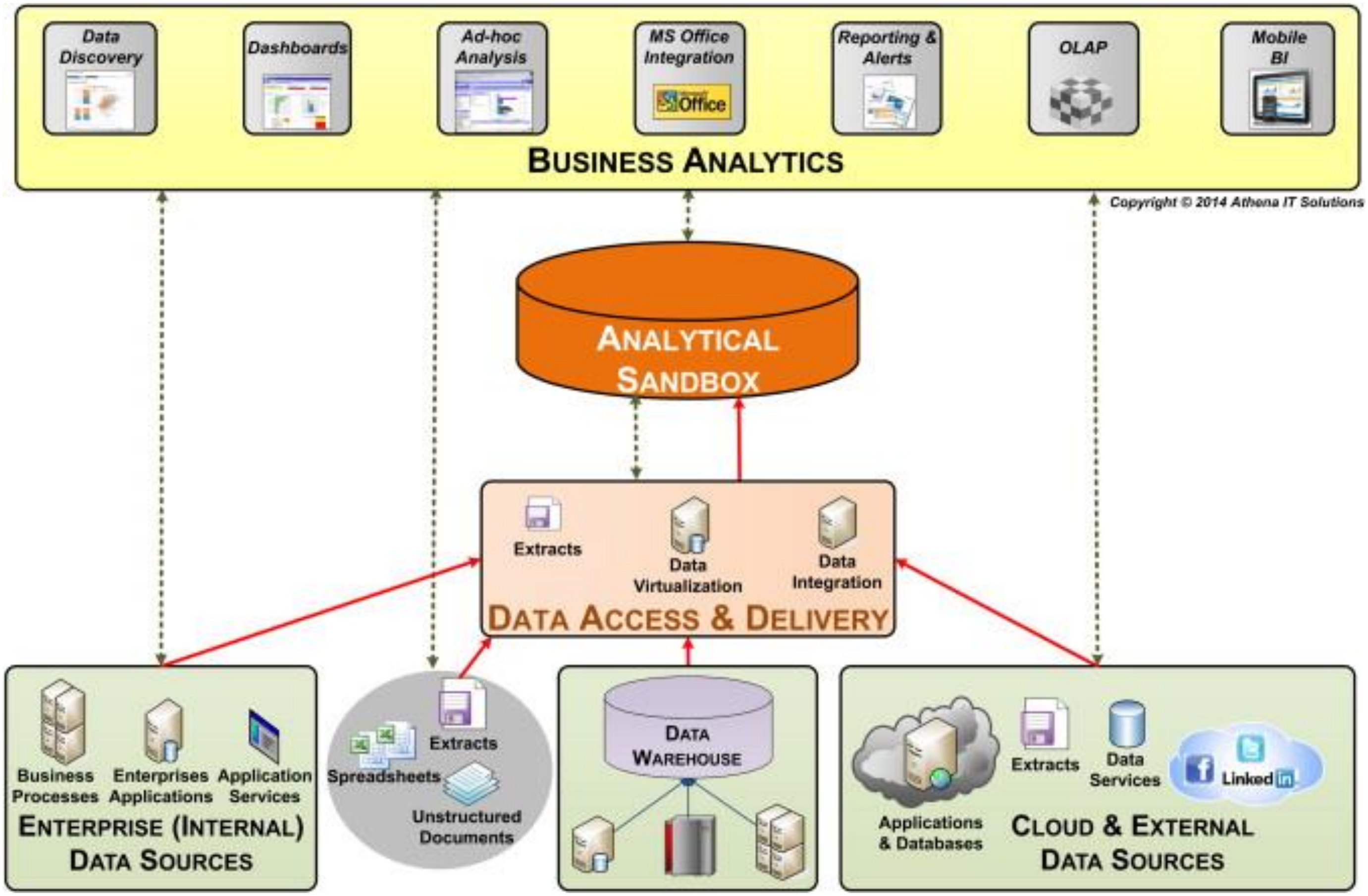
  2. Consistently leverage a database platform through a sandbox

     Enterprise Analytic Data Set (EADS)

  3. Necessary to keep scores up to date on a daily

     Embedded Scoring

Analytic Sandbox



Analytic Sandbox

- An analytic sandbox provides a set of resources with which in-depth analysis can be done to answer critical business questions. An analytic sandbox is ideal for data exploration, development of analytical processes, proof of concepts, and prototyping.
- Once things progress into ongoing, user-managed processes or production processes, then the sandbox should not be involved.
- A sandbox is going to be leveraged by a fairly small set of users.
- There will be data created within the sandbox that is segregated from the production database.
- Sandbox users will also be allowed to load data of their own for brief time periods as part of a project, even if that data is not part of the official enterprise data model.
- Data in a sandbox will have a limited shelf life. The idea isn't to build up a bunch of permanent data. During a project, build the data needed for the project. When that project is done, delete the data. If used appropriately, a sandbox has the capability to be a major driver of analytic value for an organization.

Analytic Sandbox

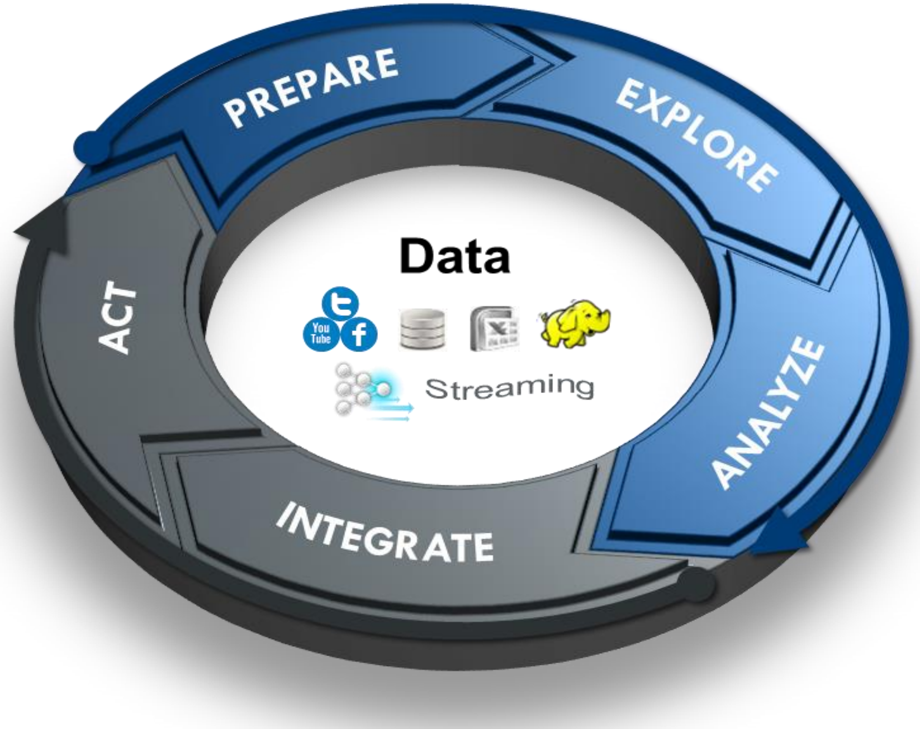
Interactive Reporting

Extend Reports with **Self-Serve Analytics & Modeling**

Integrate **Advanced Analytical Processes** within the Visualization Environment

Deployment

- Regulated
- Automated
- Governed
- Robust
- Reliable
- Decisions
- Consistent
- Documented
- Actions
- IT



Discovery

- Lots of Data
- New Data
- Experimentation
- Fail Fast
- Test & Learn
- Interactive
- Iterative
- Innovation
- Flexibility
- Data Science

The goal of an analytical sandbox is to enable business people to conduct discovery and situational analytics. This platform is targeted for business analysts and “power users” who are the go-to people that the entire business group uses when they need reporting help and answers. This target group is the analytical elite of the enterprise.

The analytical elite have been building their own makeshift sandboxes, referred to as data shadow systems or spread marts. The intent of the analytical sandbox is to provide the dedicated storage, tools and processing resources to eliminate the need for the data shadow systems.

The key components of an analytical sandbox are:

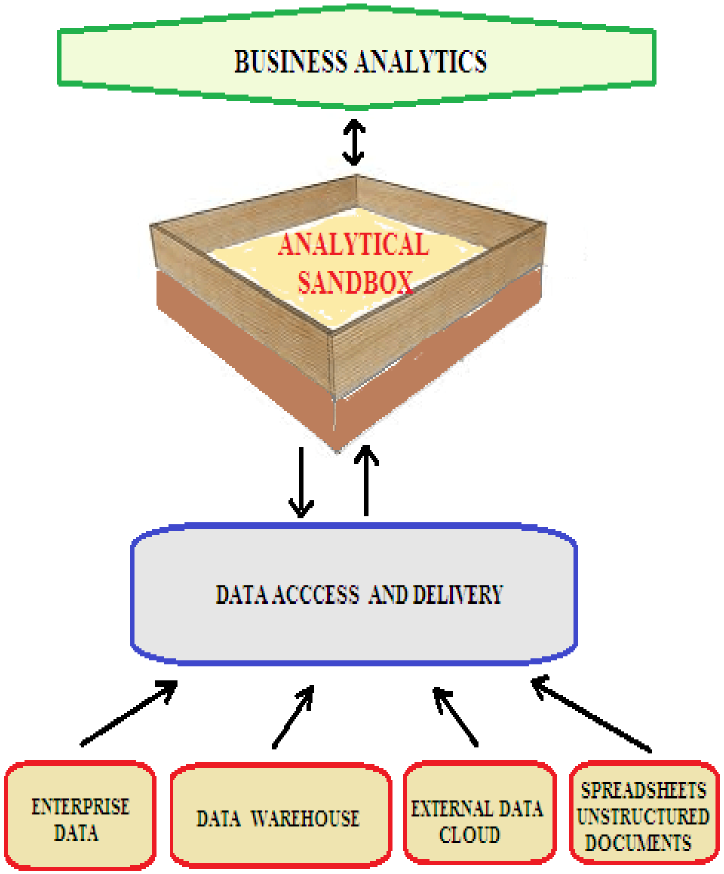
Business analytics – contains the self-service Business Intelligence tools used for discovery and situational analysis

Analytical sandbox platform – provides the processing, storage and networking capabilities

Data access and delivery – enables the gathering and integration of data from a variety of data sources and data types

Data sources – sourced from within and outside the enterprise, it can be big data (unstructured) and transactional data (structured); e.g., extracts, feeds, messages, spreadsheets and documents.

Benefits of Analytic Sandbox

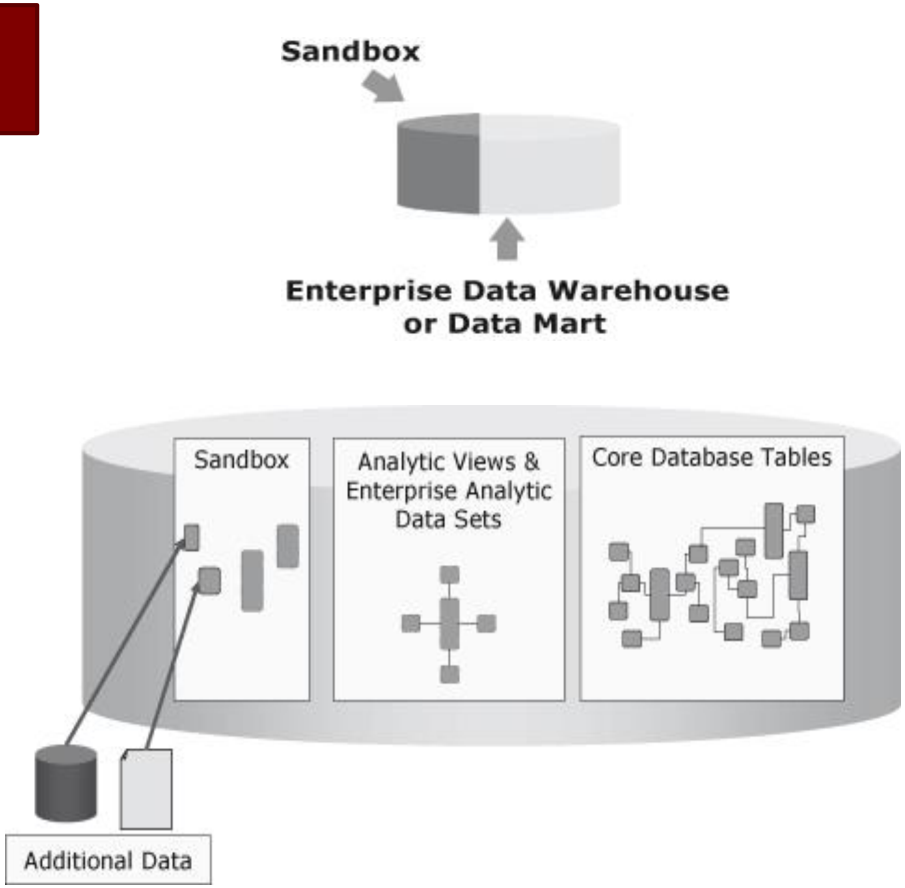


- □ ***Independence.*** Analytic professionals will be able to work independently on the database system without needing to continually go back and ask for permissions for specific projects.
- □ ***Flexibility.*** Analytic professionals will have the flexibility to use whatever business intelligence, statistical analysis, or visualization tools that they need to use.
- □ ***Efficiency.*** Analytic professionals will be able to leverage the existing enterprise data warehouse or data mart, without having to move or migrate data.
- □ ***Freedom.*** Analytic professionals can reduce focus on the administration of systems and babysitting of production processes by shifting those maintenance tasks to IT.
- □ ***Speed.*** Massive speed improvement will be realized with the move to parallel processing. This also enables rapid iteration and the ability to "fail fast" and take more risks to innovate.

Types of Analytic Sandbox

Internal Sandbox

- For an internal sandbox, a portion of an enterprise data warehouse or data mart is carved out to serve as the analytic sandbox.
- In this case, the sandbox is physically located on the production system. However, the sandbox database itself is not a part of the production database. The sandbox is a separate database container within the system.


Sandbox
Enterprise Data Warehouse or Data Mart
Sandbox
Analytic Views & Enterprise Analytic Data Sets
Core Database Tables
Additional Data

Types of Analytic Sandbox

Strengths of Internal Sandbox

- One strength of an internal sandbox is that it will leverage existing hardware resources and infrastructure already in place.
- This makes it very easy to set up. From an administration perspective, there's no difference in setting up a sandbox than in setting up any other database container on the system.
- What's different about the sandbox are some of the permissions that will be granted to its users and how it is used. Perhaps the biggest strength of an internal sandbox is the ability to directly join production data with sandbox data.
- Since all of the production data and all of the sandbox data are within the production system, it's very easy to link those sources to one another and work with all the data together.
- An internal sandbox is very cost-effective since no new hardware is needed.
- The production system is already in place. It is just being used in a new way. The elimination of any and all cross-platform data movement also lowers costs.
- The one exception is any data movement required between the database and the MapReduce environment.

Types of Analytic Sandbox

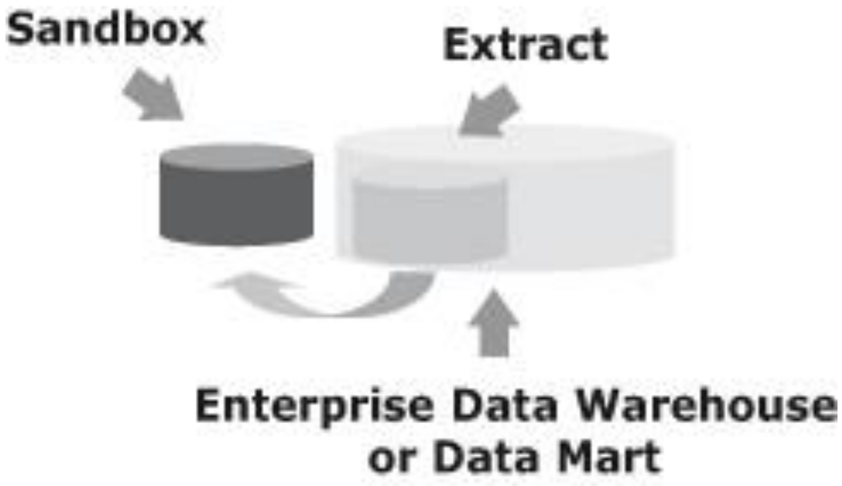
Weakness of Internal Sandbox

- There are a few weaknesses of an internal sandbox.
- One such weakness is that there will be an additional load on the existing enterprise data warehouse or data mart.
- The sandbox will use both space and CPU resources (potentially a lot of resources). Another weakness is that an internal sandbox can be constrained by production policies and procedures.
- For example, if on Monday morning virtually all the system resources are needed for Monday morning reports, sandbox users may not have many resources available to them.

Types of Analytic Sandbox

External Sandbox

- For an external sandbox, a physically separate analytic sandbox is created for testing and development of analytic processes.
- It's relatively rare to have an environment that's purely external.



Types of Analytic Sandbox

Strengths of External Sandbox

- The biggest strength of an external sandbox is its simplicity. The sandbox is a stand-alone environment, dedicated to advanced analytics development. It will have no impact on other processes, which allows for flexibility in design and usage.

- Another strength of an external sandbox is reduced workload management. When only analytic professionals are using the system, it isn't necessary to worry much about tuning and balancing.
- There will be predictable, stable performance in both the sandbox and production environments. For example, sandbox users won't have a Monday morning downgrade to their resources due to reporting needs. They'll have a steady level of access to the sandbox.

Types of Analytic Sandbox

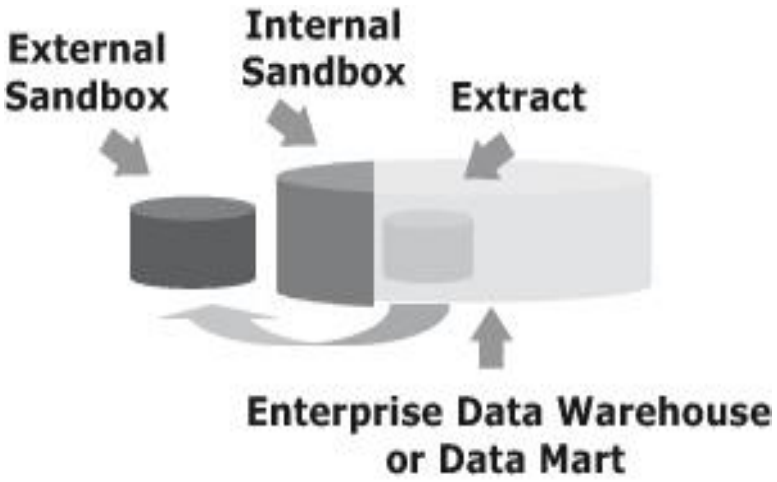
Weakness of External Sandbox

- A major weakness of an external sandbox is the additional cost of the stand-alone system that serves as the sandbox platform.
- To mitigate these costs, many organizations will take older equipment and shift it to the sandbox environment when they upgrade their production systems. This makes use of equipment that would otherwise be discarded and saves any costs associated with hardware for the sandbox.
- Another weakness is that there will be some data movement. It will be necessary to move data from the production system into the sandbox before developing a new analysis.
- The data feeds will also need to be maintained. The feeds don’t have to be too complicated, but it is an extra set of tasks to maintain and execute.
- Any data feeds should be scoped very tightly and should focus only on what is absolutely needed

Types of Analytic Sandbox

Hybrid Sandbox

- A hybrid sandbox environment is the combination of an internal sandbox and an external sandbox.
- It allows analytic professionals the flexibility to use the power of the production system when needed, but also the flexibility of the external system for deep exploration or tasks that aren't as friendly to the database.



Types of Analytic Sandbox

Strength of Hybrid Sandbox

- The strengths of a hybrid sandbox environment are similar to the strengths of the internal and external options, plus having ultimate flexibility in the approach taken for an analysis.
- It is easy to avoid production impacts during early testing if work is done on the external sandbox. When it comes time for final testing and pre-deployment work, the production sandbox can be used.
- A single MapReduce environment might augment the hybrid sandbox by supporting both the internal and external sandboxes.
- Another advantage is if an analytic process has been built and it has to be run in a "pseudo-production" mode temporarily while the full production system process is being deployed. Such processes can be run out of the internal sandbox easily.

Types of Analytic Sandbox

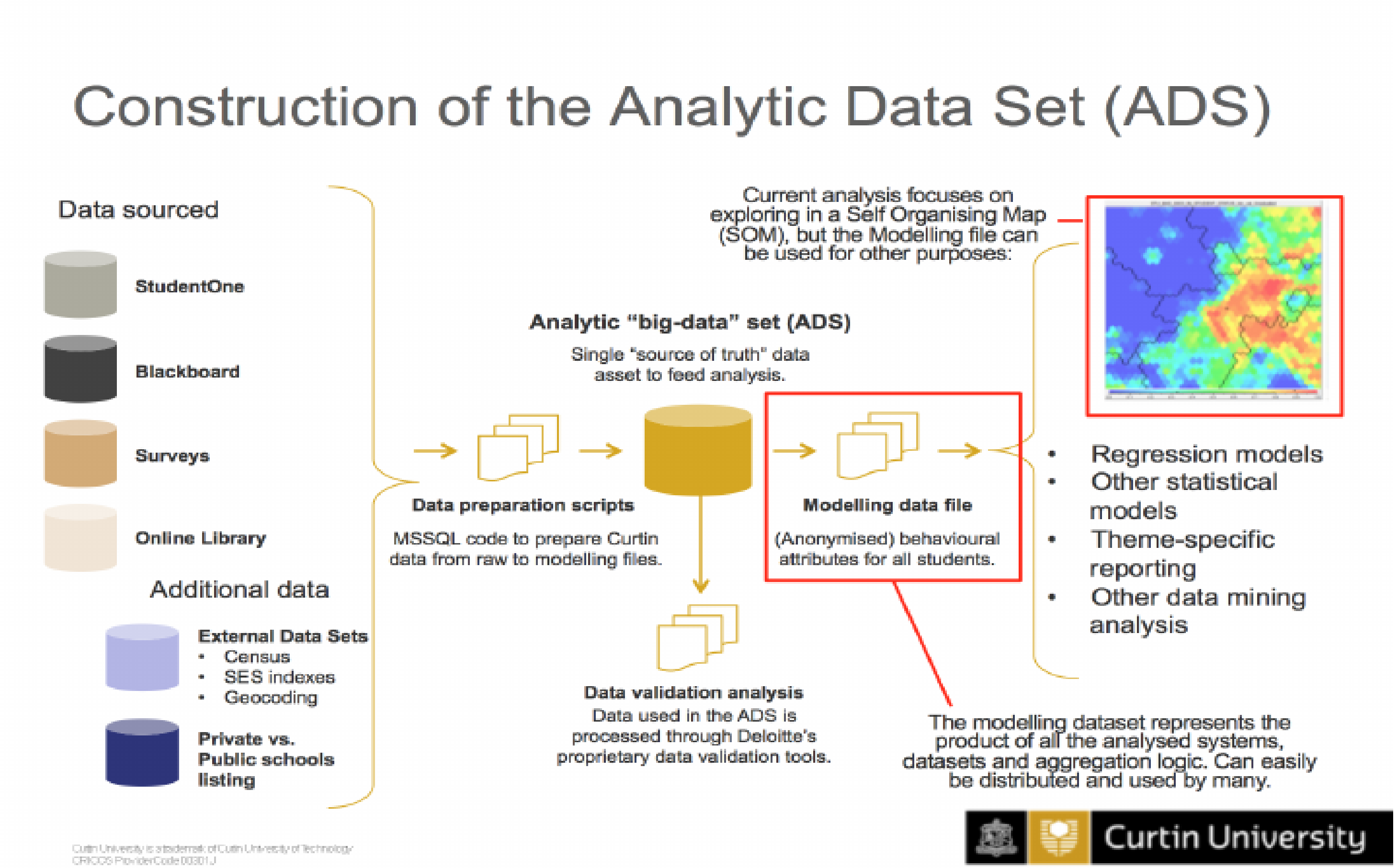
Weakness of Hybrid Sandbox

- □ The weaknesses of a hybrid environment are similar to the weaknesses of the other two options, but with a few additions. One weakness is the need to maintain both an internal and external sandbox environment.
- □ Not only will it be necessary to keep the external sandbox consistent with the production environment in this case, but the external sandbox will also need to be kept consistent with the internal sandbox.
- □ It will also be necessary to establish some guidelines on when each sandbox option is used.
- □ There ought to be certain types of activities that are earmarked for the external sandbox and certain activities earmarked for the internal sandbox.
- □ It can’t be a matter of analytic professionals arbitrarily using one or the other.
- □ The analytics team is going to have to develop guidelines and stick to them.

Analytic Data set

- An **analytic data set (ADS)** is 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 analyzed. The analytic data set helps to bridge the gap between efficient storage and ease of use.
- Most data in relational databases is stored in what is known as **third normal form.** This is a method of storing data that eliminates data redundancy but makes queries more complex.
- **Third normal form table structures are very efficient for storing and retrieving data, but they cannot be directly used for most advanced analytics efforts.**

Analytic Data set



Construction of the Analytic Data Set (ADS)

Data sourced

- StudentOne
- Blackboard
- Surveys
- Online Library

Additional data

External Data Sets
- Census
- SES indexes
- Geocoding

Private vs. Public schools listing

Data preparation scripts

MSSQL code to prepare Curtin data from raw to modelling files.

Analytic "big-data" set (ADS)

Single "source of truth" data asset to feed analysis.

Data validation analysis

Data used in the ADS is processed through Deloitte's proprietary data validation tools.

Modelling data file

(Anonymised) behavioural attributes for all students.

Current analysis focuses on exploring in a Self Organising Map (SOM), but the Modelling file can be used for other purposes:

- Regression models
- Other statistical models
- Theme-specific reporting
- Other data mining analysis

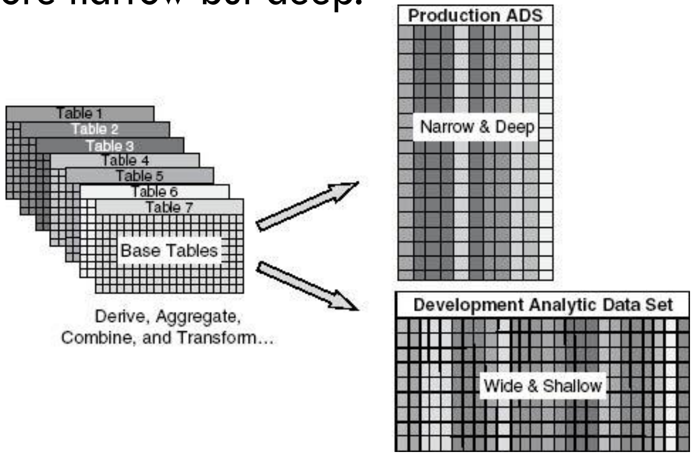
The modelling dataset represents the product of all the analysed systems, datasets and aggregation logic. Can easily be distributed and used by many.

Curtin University

Curtin University is a trademark of Curtin University of Technology
CRICOS Provider Code 00301J

Development vs Production Analytic Data Sets

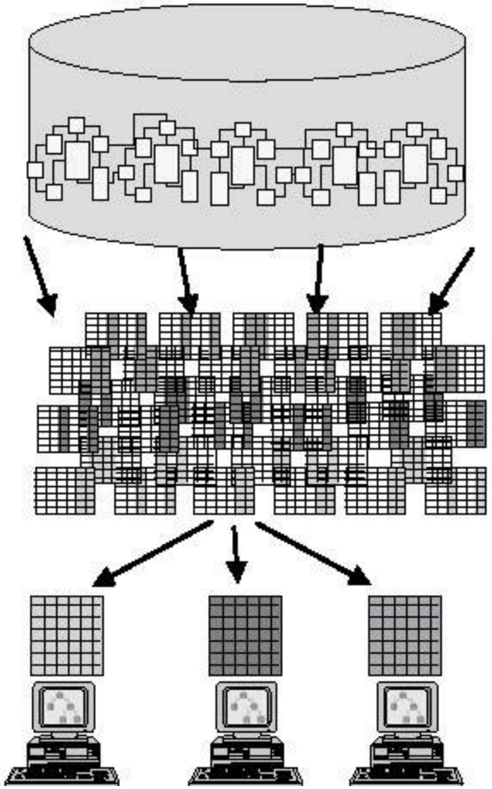
- A **production analytic data set,** however, is what is needed for scoring and deployment. It's going to contain only the specific metrics that were actually in the final solution. Typically, most processes only need a small fraction of the metrics explored during development.
- A big difference here is that the scores need to be applied to every entity, not just a sample. Every customer, every location, every product will need to be scored. Therefore, a production ADS is not going to be very wide, but it will be very deep.
- For example, when developing a customer model, an analytic professional might explore 500 candidate metrics for a sample of 100,000 customers.
- The development ADS is therefore wide but shallow. When it comes time 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

- In a traditional environment, all analytic data sets are created outside of the database.
- Each analytic professional creates his or her own analytic data sets independently. This is done by every analytic professional, which means that there are possibly hundreds of people generating their own independent views of corporate data.
- It gets worse! An ADS is usually generated from scratch for each individual project. The problem is not just that each analytic professional has a single copy the production data. Each analytic professional often makes a new ADS, and therefore a new copy of the data, for every project.

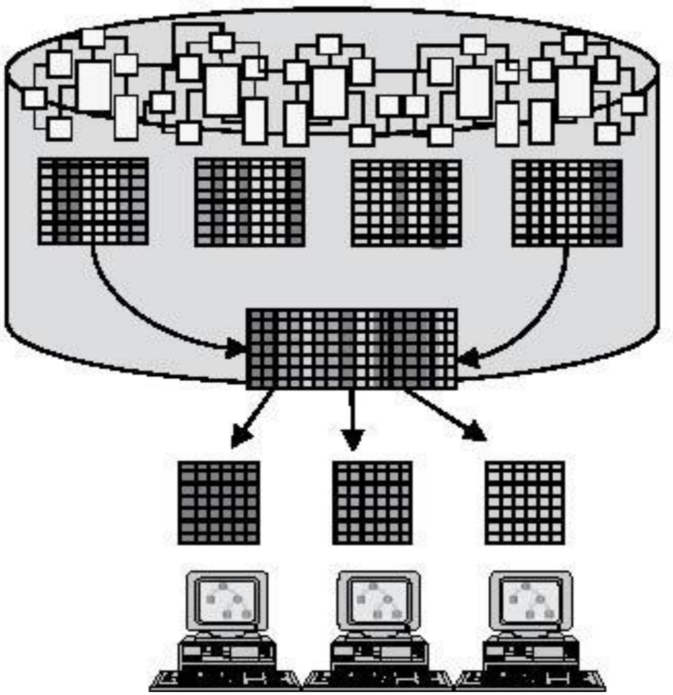
Traditional Analytic Data Set Process:
A dedicated ADS is generated outside the database for every project.



Enterprise ADS

- *An EADS is a shared and reusable set of centralized, standardized analytic data sets for use in analytics.*
- What an EADS does is to condense hundreds or thousands of variables into a handful of tables and views.
- These tables and views will be available to all analytic professionals, applications, and users. The structure of an EADS can be literally one wide table, or it may be a number of tables that can be joined together.
- An EADS is collaborative in that all of the various analytic processes can share the same, consistent set of metrics.
- An EADS is going to greatly simplify access to data by making many metrics available directly to analytic professionals without further effort. They no longer have to go and navigate the raw third normal form tables and derive all the metrics themselves. An EADS is going to greatly reduce time to results and it is a "build once, use many" endeavor.

<u>Enterprise Analytic Data Set Process:</u>
Centralized ADS tables and views
are utilized across many projects.



Embedded Scoring

- Embedded scoring involves enabling 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.
- Just to review, analytic processes often result in the outputting of a new piece of information. Examples include a customer’s likelihood to purchase a product, the optimal price point for a product, or the expected lift in sales that a specific location will see during a promotion. When the analytics developed are applied with current data, this is called scoring

Embedded Scoring

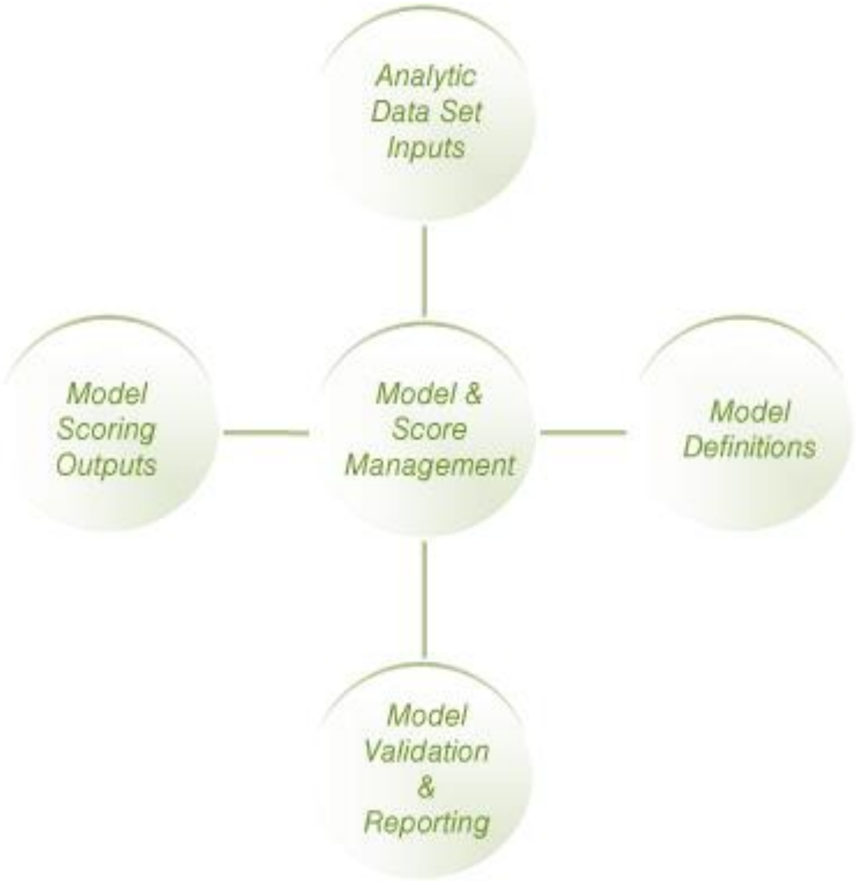
- First, scores run in batches will be available on demand. If regularly scheduled batch updates to a set of scores are done, then when a user needs to access a score it will be there waiting. It's also possible to do a batch update only when needed. For example, an organization may update the scores for the customers being added to a mail list only at the time the mail list is created.
- Next, embedded scoring enables real-time scoring. This is especially important for situations such as web offers. If someone's on a site now, he must be scored based on what is known about him right now, including what he just did on the site, to get the right offer to him when he browses the next page. Similarly, perhaps someone is on the phone with a call center. As the customer has a conversation with a call center rep, the rep inputs any new information he or she has learned. The inputting of that information might warrant an update to the customer's score so the rep knows the right path to go down next.

Embedded Scoring

- Next, embedded scoring will abstract complexity from users. It’s very easy for both individual users and applications to ask for a score. The system handles the heavy lifting. As a result, embedded scoring will make scores accessible to less technical people.
- A final benefit is having all the models contained in a centralized repository so they are all in one place. If an inventory of models and scores created is kept through a model management process, it is possible to keep track of what has been created more easily. No longer will analytic professionals across an organization keep the scoring processes they create within their specific control. Rather, they will be managed centrally and deployed for wider use.

Model and Score Management

- □ There are four primary components required to effectively manage all of the analytic processes an enterprise develops.
- □ The components include analytic data set inputs, model definitions, model validation and reporting, and model scoring output.
- □ There are commercially available tools to help with model and score management, or a custom solution can be built to address an organization’s specific needs.



Analytic Tools

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.
- It is important to note that ensemble models go beyond picking the best individual performer from a set of models. They actually combine the results of multiple models in order to get to a single, final answer.

Analytic Evolution| Ensemble

- □ The power of ensemble models stems from the fact that different techniques have different strengths and weaknesses.
- □ 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.

Analytic Evolution| Ensemble

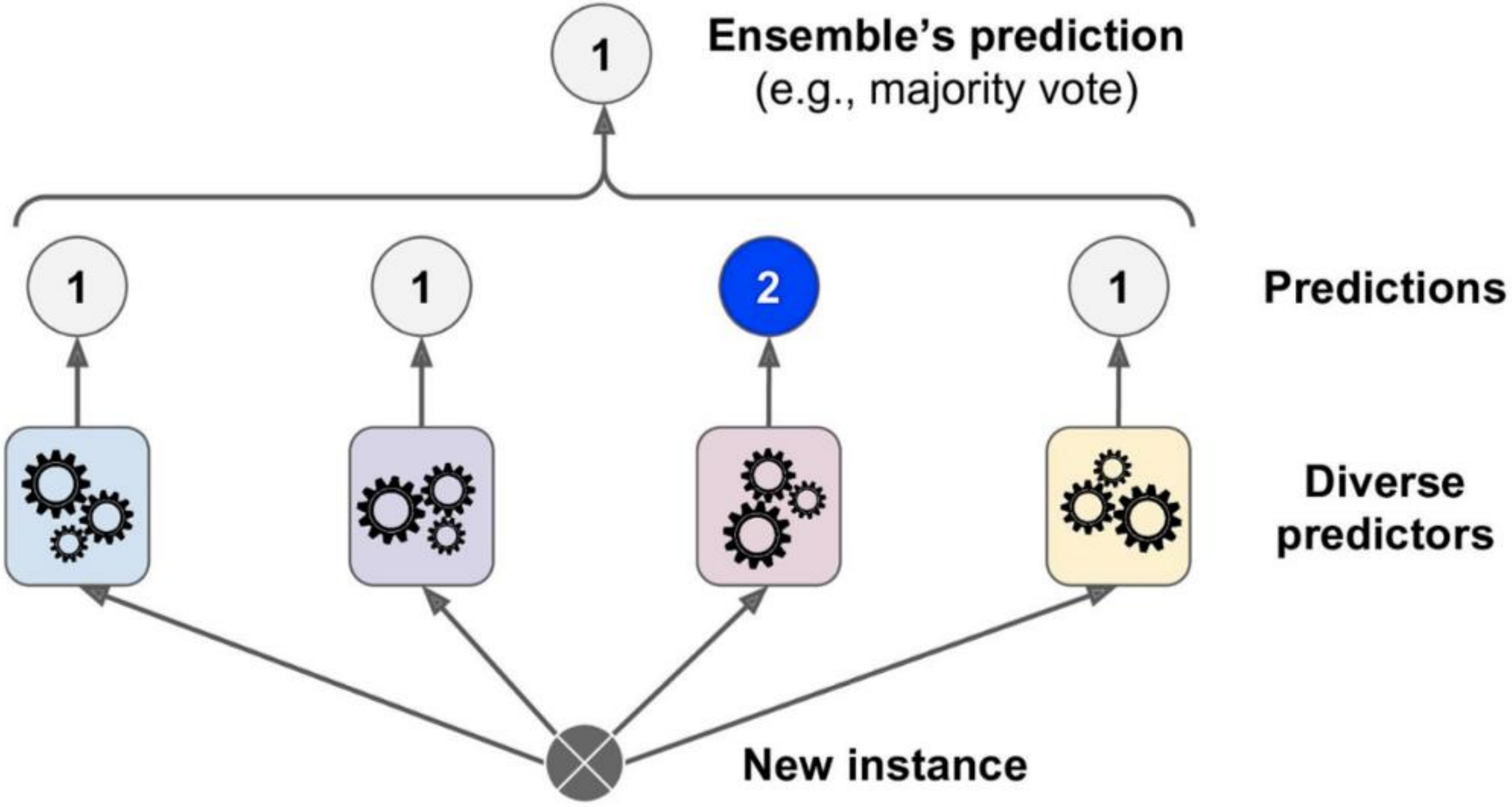
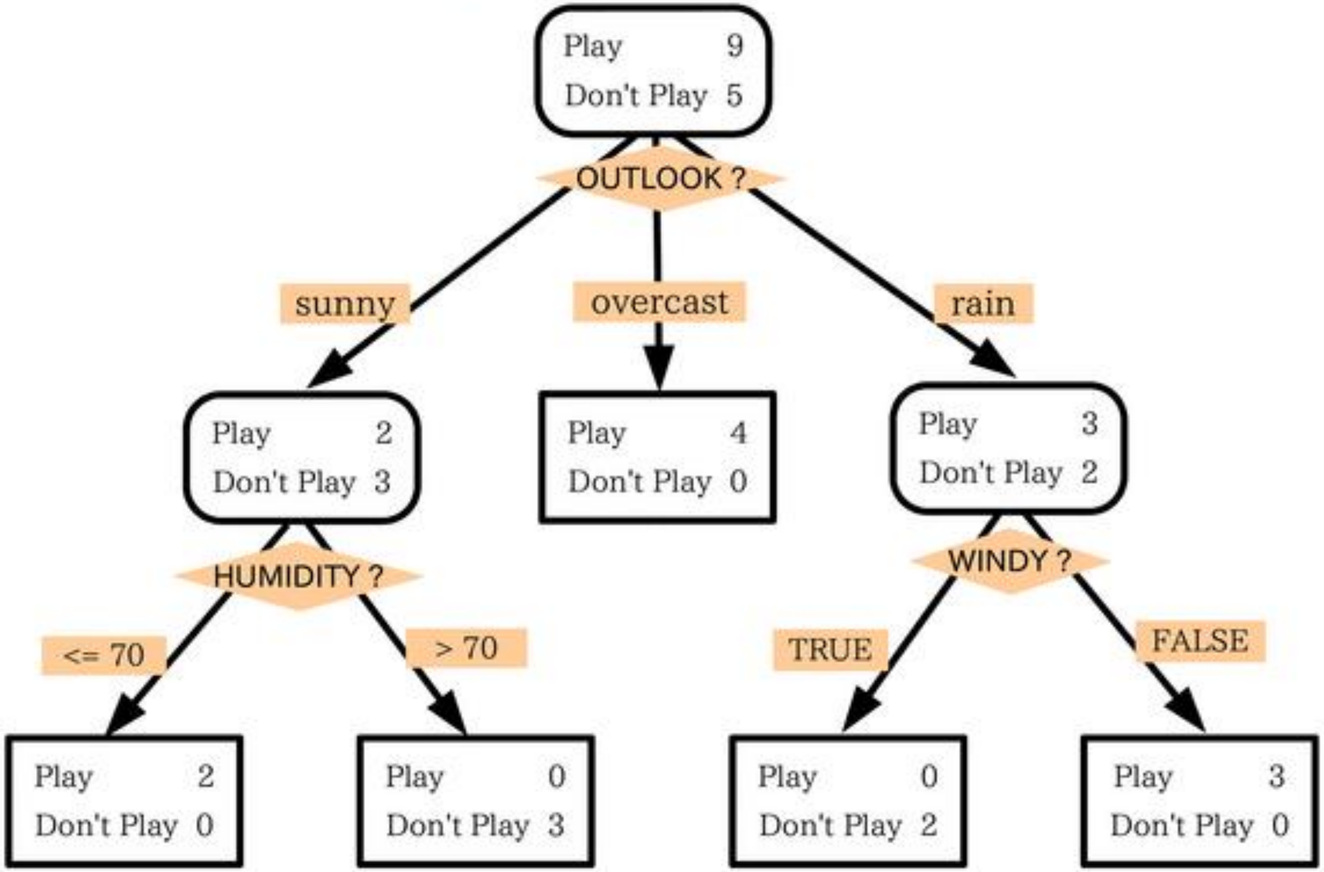


Figure 7-2. Hard voting classifier predictions

Analytic Evolution | Ensemble

Dependent variable: PLAY



Analytic Evolution| Commodity

- We’ll define a commodity model as one that has been produced rapidly and with less concern for squeezing out every ounce of lift or predictive power. Commodity models might be done via a simple stepwise analysis procedure, for example, on a mostly automated basis.
- **The goal of a commodity model 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.**
- Used appropriately, commodity models can be quite useful and can extend the **impact of analytics within an organization**. Traditionally, building models was a time-intensive task.
- As a result, it was expensive. Analysts would spend weeks or months just getting data together and then more time running models against the data. This necessitated building models sparingly and only for very high-value problems. If you had a 30- to 40-million piece mailing upcoming, then it was absolutely worth the investment to build a model. If you had an upcoming mailing of 30,000 pieces for a fairly inexpensive product, there was no way it was worth investing in a model.

Analytic Evolution| Commodity

- □ If analytic professionals are making use of a modern environment including a scalable sandbox, as well as modern processes including enterprise analytic datasets, then building models does not need to be as time-intensive as it used to be. We discussed these topics in Chapters 4 and 5. The more that standard variables are available and processing horsepower can be applied to them, the easier it is to go through the mechanics of building models.
- □ Always remember that making it easier to go through the mechanics doesn't remove the need to be diligent and to make sure the correct mechanics are done. But if a good analytics professional is driving the process, it is possible to get things done much faster.

Analytic Evolution | Text

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

Analytic Evolution | Point Solution

- 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 of point solutions include price optimization applications, fraud applications, and demand forecasting applications, among others. Point solutions built on tool suites, such as SAS, will utilize some of the generic functionality of the underlying toolset. However, the user interface will be geared specifically to a targeted set of problems. There may be many man-years of development work that go into a point solution. Organizations can consider purchasing one as an alternative to building their own solution. It can save both money and time.

Analytic Evolution| Point Solution

- Analytic point solutions have gained traction as a way to allow specific departments within an organization to utilize higher analytics in their daily business processes. These tools typically require a very high level of knowledge to install, configure, and initially set up the parameters of the analytics to be run.
- Over time, there's a lower bar for how much knowledge is required for ongoing maintenance and usage of the solution.
- This opens point solutions to a wider user base. Note that this does not violate the previous point about people not using tools if they don't understand code. Point solutions are built and configured to constrain a user to actions that are appropriate.

Analytical Tools

- SPSS - Statistical Package for the Social Sciences
- Statistical Analysis Software (SAS)
- Stata: Software for Statistics and Data Science
- *R* is a *programming language* and free software environment for statistical computing and graphics supported by the *R* Foundation for Statistical Computing.
- Tableau Software