Leveraging existing databases for advanced analytics of the future

Petroleum Network Education Conference (PNEC), May 22nd-24th 2018, Houston, Texas

Christopher Hanton, *Perigon Data Solutions, 8584 Katy Fwy, Suite 103, Houston, Texas*

James Miller, *Anadarko Petroleum, 1099 18th St, Denver, Colorado*

The increased abundance of computer power available to the oil and gas industry has moved Machine Learning based Advanced Analytics out of a niche R&D role and into mainstream modelling and predictions. These advancements have allowed for an increase in not only the possible workflows utilizing this data, but also the types and amount of data that can be processed.

The industry-wide push for this mode of data usage has created new challenges in managing the flow of data across an organization. Large amounts of data must be made available, but in a standardized and clean manner to ensure maximum efficiency and productivity from these models. Traditional linear flows of data within an organization present bottlenecks that become even more chronic when scaled to the levels of 'big-data'.

By working closely with Perigon, Anadarko Petroleum have created an enterprise database for all data pertaining to core which is accessible across the entire organization for conventional workflows, and also feeds directly into these advanced analytic models.

Background Information

In 2017, Anadarko announced the formation of its Advanced Analytics and Emerging Technologies team (AAET). The geoscience group within AAET was tasked with exploring and developing advanced workflows for making basin analysis more efficient. It was quickly realized that Anadarko's geoscience data was not managed in a way that made it easily accessible for these types of advanced workflows. Like many companies the data is maintained in "data silos" where individual assets load the data into stand-alone applications and store original documents into directory structures on the network. There is no standardization for how the data is loaded into projects or how it is structured in the storage locations. The result is there is no easy way to aggregate, format or even access the data by "big data" type applications.

After reviewing the state of affairs with AAET and Geoscience Technology management, several projects were approved within IT to begin bringing together key geologic data types into modern relational databases. The individual databases were each designed around specific data types and each needed to have the ability to feed into a "data-lake" for consumption by the AAET and asset teams. This paper specifically covers core data.

Short Term vs Long Term

The first consideration was whether to create a database of digital data with the sole purpose of storing data in a format that was consumable by analytic systems, or to create a core data repository that could manage all core data types and was easily accessible by end-user geoscientists for more conventional style workflows. Given the undeniable long-term benefits that the creation of a clean system that could

**Update: As of May 2021, "iPoint" is now "Curate"*

be easily accessed and browsed by geoscientists would provide to the organization the second option was taken. As such the focus of the project moved towards quickly and efficiently populating this database and opening up access to the Advanced Analytics teams.

The Challenge

The largest challenges faced were the considerable amounts of core data that Anadarko had and the lack of consistent machine readable formats for this data. Core reports frequently change format over time and from location to location within the same vendor. Metadata (even including UWI, Vendor Name and Report Type) is commonly held in a format which, while easy to read as a geoscientist, is not conducive to machine reading. Images and documents also make up a large amount of the stored core data and again are often missing key metadata or exist in formats that require rigorous amounts of manual work to extract.

Due in part to these problems, core data management practices in organizations of Anadarko's size generally lag behind those of other wellbore data types such as logs and formation tops. Data is often held in isolated locations and requires aggregation followed by a comprehensive QC and a standardization workflow prior to use.

This workflow is summarized in Figure 1 and results in a linear chain of dataflow.

Figure 1 Linear Data flow from conventional data management systems

This linearity results in a series of bottlenecks where, for an individual data type, each step can only be started once the previous task has been completed. This workflow also places strain on individual resources such as data techs compiling data, or those performing the QC and standardization. In the present market conditions and with a focus on efficiency, manually working through this data is no longer a viable solution.

The Solution

Given the volume and breadth of data types to consider and the number of potential sources, a more intelligent and rounded approach was required.

Leveraging technology and experience from similar implementations and combining it with Anadarko's own data knowledge and aims, Perigon created a system of drop-box enabled autoloaders that ran a range of pre-defined business rules on the data prior to loading into their wellbore data management tool, iPoint. Once the data was loaded into iPoint, accessibility to both end users and the advanced analytic processes was directly provided. Figure 2 gives a schematic overview of the final system alongside a detailed description of the stages below.

**Update: As of May 2021, "iPoint" is now "Curate"*

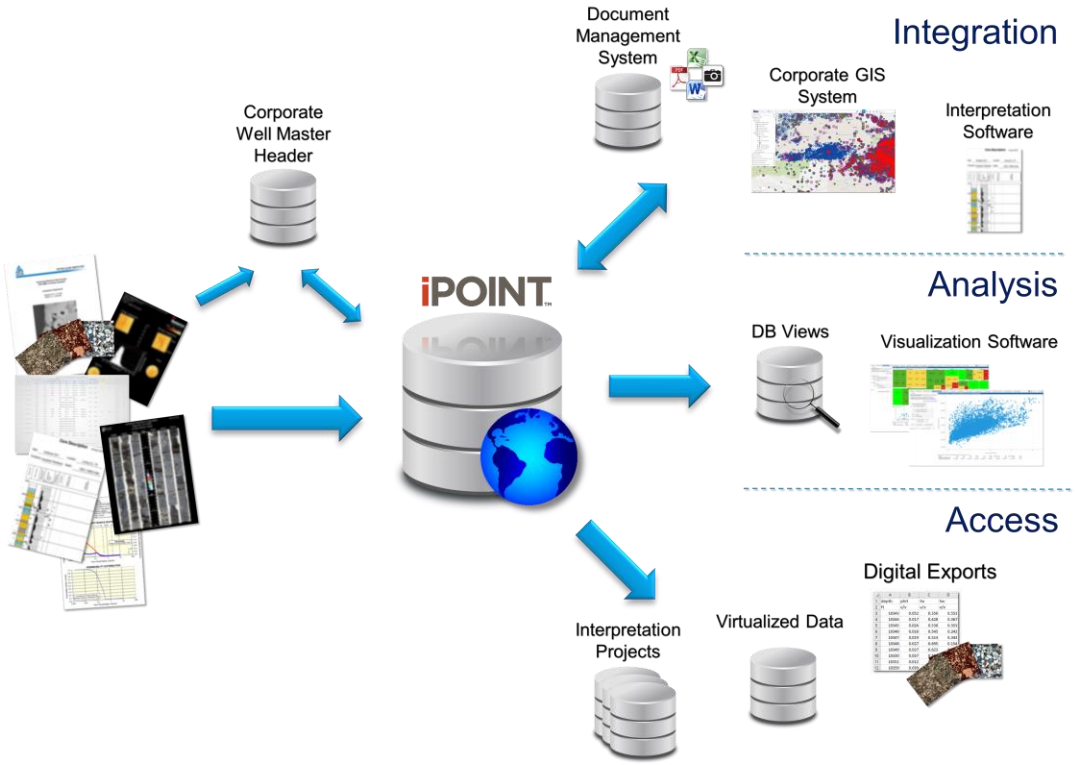

Figure 2 Schematic of final implementation

1. Data Manipulation: Given the lack of standards for core report formats and inconsistent placement of key metadata in files, a certain amount of data massage is required prior to loading. This consisted primarily of identifying the UWI for the data sources and appending it to the beginning of the filename. Image files also had additional metadata (where available) added to their titles indicating image type and resolution/scale.
2. Aggregation and Loading from Multiple Sources: The creation of drop-boxes in a shared environment removes the need for a single dedicated resource to load to the system. Users require no training above being able to drag files into a folder, speeding up the loading process and opening it to a wider potential audience. This also increases the likelihood of undocumented legacy data that may be on hard-drives or isolated file systems being loaded.
   An illustrative example of a folder structure is shown in Table 1. By having folders split up on coarse disciplines (as opposed to a separate folder per report type) it reduces the amount of knowledge the data loader requires (e.g. just to know that a file is conventional core analysis, not the preparation method, sample type etc.).

**Update: As of May 2021, "iPoint" is now "Curate"*

Folder Level 1	Folder Level 2	Reports
Vendor 1	CCA	Core Plug Data
		Whole Core Data
		Dean-Stark Data
		Bulk Density
		Sidewall analysis
	Rock Pyrolosis	Leco TOC
		PyroChromatography
		Pyrograms
		Pyrolysis and Oxidation
		Visual Kerogen Analysis
		VRo
Vendor 2	CCA	Core Plug Data
		Whole Core Data
		Bulk Density
		Sidewall analysis
	Rock Pyrolosis	Leco TOC
		PyroChromatography
		Pyrolysis and Oxidation
		Visual Kerogen Analysis
		VRo
		RockEval

Table 1 Example of implemented folder structure for drop-boxes

3. Application of Business Rules: Business rules and standards are created prior to loading, based on the availability of the data and an understanding of how the data will be used by end-users and advanced processes. These are then automatically applied to each data type upon load, along with other rules such as the management of data duplication. Well header information is retrieved for new wells automatically from the master data management solution, ensuring correct values are stored.
4. Load/Quarantine: The loading process runs automatically though all data in the folders. Data which is correctly formatted and meets all business rules is loaded, while nonconforming data is moved to the quarantine folder. Upon processing, email notifications are delivered to relevant parties, providing updates and the status of data loads
5. Access via Web-Based Application or Direct Query via Tools: iPointWeb provides site wide access to all authorized users within the Anadarko organization, allowing querying, viewing and exporting of all datatypes quickly from any authorized web-enabled device. Database 'views' are made available for the advanced analytic tools to access via SQL queries, allowing the AAET team to choose which data they require for their workflows.

**Update: As of May 2021, "iPoint" is now "Curate"*

Conclusion

After the creation and roll out of this system, Anadarko were able to utilize a single database for both the shorter term goal of providing data to their Advanced Analytics team and also as a long term solution to getting core data into the hands of subsurface engineers.

The autoloading drop-box approach enables anyone with access to the area and a basic understanding of data types to add to the system, ensuring the maximum amount of data is retrieved and placed into the system.

In this example a new database was created for the project, however the workflow for exposing data within a relational database to advanced analytic tools is applicable to any legacy database.

A final recommendation from this report would be for vendors of non-standard data (such as core, geochem, PVT etc.) to work with clients to produce reports that are natively machine readable by default. By understanding the processes which the reports are put through by the client in order to extract and collate the information, vendors could create standardized report formats that eliminates the need for pre-loading data manipulation.

**Update: As of May 2021, "iPoint" is now "Curate"*