IDC PERSPECTIVE

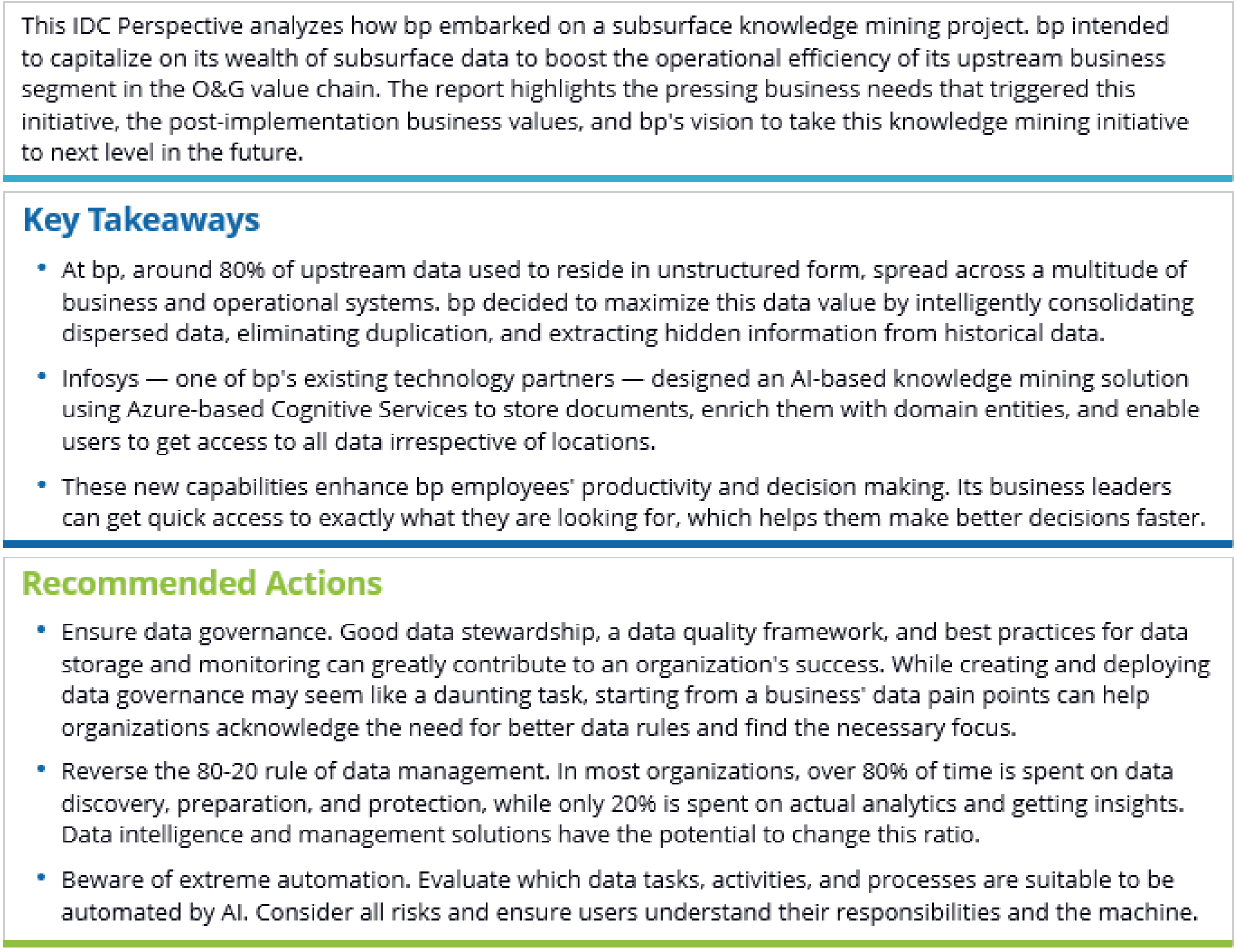
Mastering Unstructured Subsurface Data Management: bp's Knowledge Mining Project

Gaurav Verma

EXECUTIVE SNAPSHOT

FIGURE 1

Executive Snapshot: Mastering Unstructured Subsurface Data Management: bp's Knowledge Mining Project



This IDC Perspective analyzes how bp embarked on a subsurface knowledge mining project. bp intended to capitalize on its wealth of subsurface data to boost the operational efficiency of its upstream business segment in the O&G value chain. The report highlights the pressing business needs that triggered this initiative, the post-implementation business values, and bp's vision to take this knowledge mining initiative to next level in the future.

Key Takeaways

- At bp, around 80% of upstream data used to reside in unstructured form, spread across a multitude of business and operational systems. bp decided to maximize this data value by intelligently consolidating dispersed data, eliminating duplication, and extracting hidden information from historical data.
- Infosys — one of bp's existing technology partners — designed an AI-based knowledge mining solution using Azure-based Cognitive Services to store documents, enrich them with domain entities, and enable users to get access to all data irrespective of locations.
- These new capabilities enhance bp employees' productivity and decision making. Its business leaders can get quick access to exactly what they are looking for, which helps them make better decisions faster.

Recommended Actions

- Ensure data governance. Good data stewardship, a data quality framework, and best practices for data storage and monitoring can greatly contribute to an organization's success. While creating and deploying data governance may seem like a daunting task, starting from a business' data pain points can help organizations acknowledge the need for better data rules and find the necessary focus.
- Reverse the 80-20 rule of data management. In most organizations, over 80% of time is spent on data discovery, preparation, and protection, while only 20% is spent on actual analytics and getting insights. Data intelligence and management solutions have the potential to change this ratio.
- Beware of extreme automation. Evaluate which data tasks, activities, and processes are suitable to be automated by AI. Consider all risks and ensure users understand their responsibilities and the machine.

Source: IDC, 2021

SITUATION OVERVIEW

This IDC Energy Insights case study focuses on bp (a global oil supermajor). In 2018, it launched an ambitious project to make all unstructured subsurface data accessible across its upstream operations. The project is part of bp's larger multiyear business information management program called "Document Neighbourhood Project."

IDC interviewed Tracey Pearce, Senior Subsurface and Knowledge Management Specialist at bp, on the major steps of the development and implementation of the knowledge mining solution that sits at the core of the project. Building on a strategic partnership with Infosys, Microsoft, and Sword Venture, bp's subsurface data management project has been a great success. The company can now make all data available to all relevant employees, irrespective of locations, for better decision making.

IDC Energy Insights' Case Studies Series

IDC Energy Insights' case study series provides oil and gas (O&G) companies with fact-based, consistent, and independent views on interesting projects implemented across the world. They focus on digital transformation (DX) initiatives, IT and operational technology (OT) solutions implementations, and more broadly, energy technology initiatives that contribute to efficiency, innovation, and sustainability. Collaborating with O&G companies and technology providers' personnel directly involved in such projects, IDC Energy Insights analysts gather all relevant information and analyze the approaches taken and the solutions' success in meeting stated goals.

Company Overview

bp Plc is a multinational, vertically integrated British O&G company that employs around 70,000 employees worldwide. bp traces its origins back to the Anglo-Persian Oil Company that discovered oil in Iran in 1908, and today it is one of the top 3 non-state-owned oil supermajors. Besides its leadership position across most of the O&G value chain – exploration and production (E&P), refining, and fuel retail – bp also invests in and supplies energy from low-carbon and renewable sources and is fully committed to growing its sustainable energy business. In fact, the company was the first oil supermajor to expand its activities beyond fossil fuels in the early 2000s, recently announcing its ambitious target of becoming a net-zero emitter from its own operations by 2050.

Unstructured Data in the Upstream Business

A vast amount of structured and unstructured data is created daily in the E&P business. While all oil companies have developed processes and systems to manage the structured portion of this data, unstructured data is more difficult to handle and often remains scattered across multiple locations – in employees' personal computers, emails, or as hard copies. They can contain information about anything: drilling reports, borehole logs, leasing agreements, third-party obligations, farm-in/farm-out deals with peers, images of geological basins, historical survey charts, seismic interpretation of lithology, facies, etc. There is a vast amount of valuable and reusable information stored in unstructured data that it is not trivial to extract at scale.

The Approach

Business Needs and Project Objectives: How the Idea of Knowledge Mining was Born at bp

Information about the subsurface is very costly to generate, and it remains highly relevant to the upstream business for a long time. For many upstream business decisions, bp's business leaders need to refer to information gathered historically, sometimes dating back several decades. The company operates in most major oilfields and works with thousands of contractors and vendors, generating massive amounts of data daily across its subsidiaries, site offices, and field operations all over the world. Around 80% of this data resides in unstructured form, spread across a multitude of business and operational systems.

At the group level, bp has two datacenters providing data services to each of the Eastern and Western Hemispheres. All data from operations in the Eastern Hemisphere typically resides and is accessible within the eastern data service and vice versa. As a result of this setup, employees often didn't have any immediate means of accessing information residing in the opposite geographical data silo. They often spent more time looking for the required information than working on it, most of the time with unsatisfactory results, as not all data was discoverable. For example, bp had multiple libraries in different disks in the eastern datacenter that most employees had no clue of. People operating in the Eastern Hemisphere didn't know how many such libraries existed in the western datacenter, and no one really knew how many existed in total. This used to cause significant inefficiencies in some of bp's core upstream business processes. Besides, the company was losing information because when people left the organization, knowledge often left with them.

This is how the idea of bringing all file libraries in one place was born. The initiative was called the Document Neighbourhood Project and was about "finding a home for the documents" so that users knew where to keep unstructured data.

With upstream being one of the most cost-intensive parts of the value chain, bp decided to make the program even more valuable for this business segment and made its information management team work on unstructured subsurface data as a specific extension to the Document Neighbourhood Project. The objective was to maximize data value by intelligently consolidating dispersed data, eliminating all duplication, and extracting hidden information from historical data. While it may look straightforward on the surface, organizing unstructured data to make information intuitively discoverable to all bp's upstream employees, regardless of their location, was far from easy.

Selecting the Solution: The Foundation Stone

Selecting the solution was one of the first challenges as there was no off-the-shelf solution that bp's document control and information management team could have simply implemented. With the business needs from operations, the group turned to Infosys — one of bp's existing technology partners. Infosys' deep understanding of bp's landscape, owing to a longstanding relationship with the company and its proven delivery capability with advanced cloud solutions, is what led bp to select Infosys for this program.

Infosys worked closely with bp's internal stakeholders to further refine the business requirements and set up a solution road map. Its developers and solution architects were challenged to propose the best solution for a massive unstructured data store of more than 75TB. The core need was to develop a tool that could bring forth the wealth of subsurface domain information hidden into non-standard storage artifacts. For that, Infosys proposed using machine learning and natural language processing techniques to efficiently tag upstream business entities.

To develop an innovative, scalable, and accessible solution to store significant amounts of data of disparate nature, the project team soon realized a cloud platform was a first critical technical requirement. To that end, Infosys engaged Microsoft and brought in a knowledge mining solution built on Azure Cognitive Services and Cognitive Search.

Solution Description

Based on Microsoft's knowledge mining framework, bp's solution architecture was built on an Ingest-Enrich-Explore model.

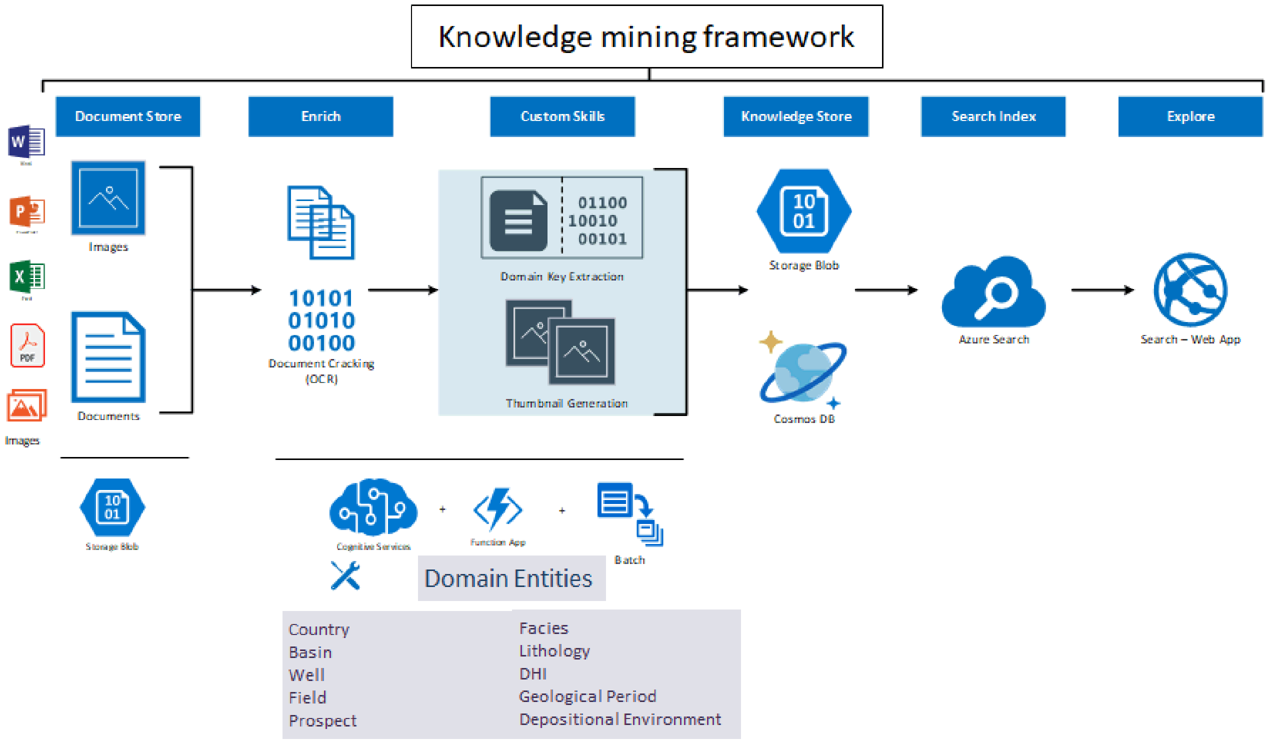
Ingest: Breaking Down Silos

The first phase of the project was called "document discovery." It involved finding all documents containing unstructured data relating to the upstream business and moving them into Blob (object) storage containers on the Microsoft Azure cloud to what was called the "Document Store." It was a humongous task. "The biggest hurdle in this process was the migration of the enormous amount of data to the cloud. That alone, could have easily taken over six months," said Pearce.

Infosys' developers and architects were challenged to find the right tool to accelerate the process, and after evaluating several candidates, they eventually picked Microsoft's Fast Data Transfer as the option that ensured the fastest possible migration.

FIGURE 2

bp's Knowledge Mining Solution Architecture



Source: Infosys, 2021

Enrich: Making Use of Every Byte of Domain Data

bp intended to create a domain-specific intelligent search platform that enabled its business leaders to easily access all past business decisions and related key operational information. With this business need in mind, the team realized that they would have to extract the richest possible set of metadata from subsurface documents, which meant capturing every single domain entity (e.g., field, well, facies, lithology, geological period) and attribute that bp users would want to filter, sort, and group the information by.

The documents and images that had been moved to the Document Store were enriched using Azure Cognitive Search built-in functionality including document cracking (i.e., extraction or creation of text from non-text sources) and skills such as natural language and image processing (e.g., entity and optical character recognition, language detection, key phrase extraction). Furthermore, custom skills were developed and added to the pipeline to extract business domain entities and other relevant document properties (e.g., author, size, date of creation, and modification), and Azure Batch was used to scale the actual computing. All extracted contents and metadata were transferred into a new storage Blob called "Knowledge Store" for further analysis.

In this phase, extracting domain-specific entities and metadata from some of the largest image files presented a significant challenge that even the solution's built-in processing skills struggled with. To address this, Infosys' developers created a series of enrichment algorithms that used a combination of image enhancement techniques to achieve higher accuracy.

Explore: Enforcing a Culture of Data-Driven Decision Making

Enabling users to navigate through a wealth of domain data with maximum ease was a of the focus of the bp information management team. With documents enriched with domain entities and metadata now available in the Knowledge Store, the next step was to build the smart search application to easily explore the subsurface information via keyword or geospatial search.

The solution development team configured Azure Cognitive Search on the extracted metadata and domain entities to index them. Infosys developers built a web search application for the indexed contents to make them searchable. Finally, to enable map-based spatial search, the team integrated the search application with bp's in-house geospatial platform – OneMap – leveraging dedicated GIS consulting services from Sword Venture.

The user interface of bp's subsurface knowledge mining system was called Upstream Subsurface Library (USL). USL enables users retrieve all the basin and related geological, geophysical, operational data, field information, well-drilling information, and many more relevant entities within a geographic area by simply drawing a polygon on its map interface. They can then refine their search by filtering out various business entities based on keywords.

Solution Deployment

From Kick-Off to a Working Prototype in Six Months

As mentioned, the urgency of utilizing the wealth of unstructured data in bp's upstream decision-making process was the primary instigator of the project. It was April 2018 when the bp information management team took up the challenge to work closely with the line of business (LOB) to solve the company's chronic lack of unified information management, and the project was kicked off.

Cross-functional collaboration and the inclusive work environment facilitated by bp enabled the solution team to work in an agile fashion and rapidly develop a prototype. This was achieved by a team of bp explorers, information managers, along with Infosys cloud architects to model a truly scalable and intelligent cloud solution, the first of its kind on many aspects. By October 2018, the team had managed to:

- Migrate files to the storage Blob
- Extract domain entities
- Define metadata sets
- Develop the script for auto-extraction of metadata
- Georeference all files and extracted information
- Integrate into bp's OneMap geospatial platform

The final product underwent several improvements and refinements through a series of trials that took place during the following eight months. The business-critical nature of this data meant bp placed utmost emphasis on the cybersecurity aspect of the project. This held back the solution's final rollout and even caused its approval to be put on hold for some time. After a thorough review process, additional security layers were applied to bp's data into the Blob storage and Cosmos database, which enabled the solution to go into production in June 2019.

Business Value

Boosting Productivity and Making Data-Driven Decisions a Reality

The deployment of the knowledge mining tool developed by bp's information management team has brought a paradigm shift in the way the company's employees access business-critical information, multiplying productivity and decision-making quality.

With all subsurface business information only a few clicks away within a single cloud library, users can finally spend more time analyzing data than looking for it.

Business leaders, in turn, get quick access to exactly what they are looking for, helping them make better decisions faster. Exploring a new block, entering a joint venture for a new oilfield, and approving or rejecting oil wells are examples of the type of business-critical and capital-intensive business decisions upstream business leaders are required to make. bp estimates that better-informed decision making enabled by the knowledge mining tool will provide an estimated $50 million to $250 million in value.

Lessons Learned

The Unconventional Approach That Put bp Ahead of the Curve

Several factors contributed to the project's success. One of the most important was that the information management team owned the business project, rather than the IT department leading the program. This ensured not only LOB's commitment, but also that the right mix of business expertise was available to the team. It also helped foster a shared vision between internal and external stakeholders, including subsurface operations executives, the information management team, and technology partners, which was also a critical success factor.

The use of DevOps and agile also contributed to the project's success. The information management team involved Infosys in bp's internal business meetings through regular touchpoints. A DevOps team was assembled, blending information management and document control personnel with solution architects and developers. The active involvement of the LOB fostered a deeper understanding of business needs and desired outcomes on the part of developers and solution architects. The agile methodology enabled the DevOps team to work closely, meet often and brainstorm, fail fast and learn faster, and do things at scale. Ultimately, this helped speed up development considerably, with tasks that could have easily taken up to six months executed in a matter of weeks. "DevOps doesn't only mean an agile methodology, but also bringing dev and ops together in the fullest sense," said Pearce.

Next Steps: Raising the Bar

Now that the standard is set, bp wants to raise the bar by making information management a true enabler for the business. A lot is already in the pipeline:

- **Ongoing refinement of the knowledge mining model.** With bp re-inventing itself resulting in new business entities, there is an immediate need to re-align the model with the new organizational structure. The solution needs to transform into an enterprisewide data discovery toolkit. With metadata auto-extraction and indexing capabilities in place, more sets of data can be ingested to further enrich the domain information that bp's business entities have access to. This will build a richer knowledge store of searchable data over time.
- **Development of a knowledge graph.** A potential future development of the currently deployed solution is the creation of an enterprise knowledge graph. This would enable bp users to be presented with search results in a more graphically rich form, helping them navigate through complex interconnected information.
- **Reaching beyond unstructured data.** bp has a forward-looking data strategy whose goal is not only to organize and integrate data, but also to extract the latent value in it (for instance, by

extracting and reorganizing historical and current well-log data, as well as running analytics using AI and ML capabilities to obtain more accurate business insights).

- **Improving structured databases.** In the long run, another focus in which bp is actively investing is to make structured databases smarter through innovative technologies such as cloud, analytics, and automation.

ADVICE FOR THE TECHNOLOGY BUYER

- **Ensure data governance for the win.** Many O&G organizations find data governance very challenging, and not many have successfully created an effective governance model. Moreover, this is one of the factors that hold O&G companies back in their DX journey. Good data stewardship, a data quality framework, and best practices for data storage and monitoring can greatly contribute to an organization's success. While creating and deploying data governance may seem a daunting task, starting from a business' data pain points can help organizations acknowledge the need for better data rules and find the necessary focus.
- **Reverse the 80-20 rule of data management.** In most organizations, over 80% of time is spent on data discovery, preparation, and protection, while only 20% is spent on actual analytics and getting to insights. Often, data management stops at the data discovery stage because people can't find what they are looking for. Data intelligence and management solutions based on intelligent technologies have the potential to change this ratio by not only giving users the ability to find data more easily, but also enabling them to understand the detailed context of the data. The maturity and availability of these technologies gives organizations an opportunity to quickly experiment with smaller data projects and eventually define a platform and a set of best practices that can be reused.
- **Beware of extreme automation.** Evaluate which data tasks, activities, and processes are suitable to be automated by AI functionality. Ensure that you have considered all the risks and that users understand their responsibilities and that of the machine. Even with automation, humans should be able to understand and explain outcomes. Most AI-based automation involved use mathematical algorithms for modeling and prediction, and its interpretation of reality should be continually tested by humans.

LEARN MORE

Related Research

- *Impact of IT-OT Integration on Oil and Gas Operations* (IDC #EUR147433821, February 2021)
- *Oil and Gas Industry Quarterly Update: October-December 2020* (IDC #EUR145815521, January 2021)
- *IDC MarketScape: Worldwide Oil and Gas Asset Performance Management 2020-2021 Vendor Assessment* (IDC #EUR147032820, December 2020)
- *IT-OT Integration Across the European Oil and Gas Industry: How We're Doing* (IDC #EUR147006120, November 2020)
- *IDC FutureScape: Worldwide Oil and Gas 2021 Predictions* (IDC #US45818220, October 2020)
- *Oil and Gas Industry Quarterly Update: July-September 2020* (IDC #EUR145815420, October 2020)

Synopsis

This IDC Perspective analyzes how bp embarked on a subsurface knowledge mining project. bp intended to capitalize on the wealth of subsurface data it built over the years to boost the operational efficiency of its upstream business. The report highlights the pressing business needs that triggered this initiative, the business value provided by the project, and bp's ambition to take this knowledge mining initiative to the next level in the future.

"A recent fad for upstream data search platforms is pushing oil companies to leverage innovative technologies such as AI, cloud, and Big Data analytics," said Gaurav Verma, research manager, IDC Energy Insights. "While developing a search engine for structured data is relatively easy for large organizations, getting value from unstructured data is a very complex endeavor. bp's knowledge mining solution — co-developed with Infosys and Microsoft — is capable of extracting domain entities from historic unstructured upstream data, something that many other oil companies are still experimenting with."

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

IDC Italy

Viale Monza, 14
20127 Milan, Italy
+39.02.28457.1
Twitter: @IDCitaly
idc-insights-community.com
www.idcitalia.com