Support Services
Directorate

KENT POLICE ESSEX POLICE

Supporting policing
in Kent and Essex

IT SERVICES
Data Transformation Project – TMT Briefing Paper

October 2024

1. The Data Transformation Project

The project seeks to centralise and modernise the existing data reporting landscape and toolset whilst upgrading Management Information and Business Intelligence reporting from Business Objects to Power BI.

Complexities and rigidity of Business Objects result in considerable developmental resource requirements for regularly used reporting. Kent & Essex Police are ready to make use of cloud technology to implement the latest in data transformation processes, storage and retrieval.

Having adopted Microsoft Azure as the cloud platform of choice there will be significant re-engineering needed around the way data is pulled from source databases and data systems, into an environment suited specifically to reporting. There is considerable business logic already written in Business Objects that will be carried forward but will need to be re-engineered to run in the Azure ecosystem. This is where most of the effort will be required in order to maintain the best of what is currently available, in the new environment.

The design of the ‘to-be’ architecture will utilise best practices for access-controlled deployment, alongside innovative governance tools that encapsulate the data journey from source through to output. These will be aligned with robust security that meets or exceeds current obligations.

As newer Business Intelligence tools such as Power BI have become mainstream and evolved into enterprise grade offerings, their ability to integrate with modules of the cloud ecosystem in which they operate is where significant benefits can be realised, that make the transition from previous generation systems such as Business Objects worth the effort.

2. Background

The current data reporting environment is driven by Business Objects. Over the c. 20 years it has been in operation, the number of data sources and systems it connects to has steadily grown to around 50 that currently serve approximately 53,000 reports ! Many of the reports connect to the same subset of data and are versions of the original report that have evolved over time. In some cases the versions simply differ from their originals only slightly, for example by a different date range.

The changes to existing reports as well as creation of new reports from Business Objects falls to the Software Development Team who have created many Universes in Business Objects with associated custom calculations and aggregations, readied for specific report types, data extracts and end users.

Centralised data that is logically related, which allows comparison and correlation between source systems' data, is what will be created in Azure, for this project. With Business Objects being retired, the transition into Azure allows for some architectural alterations that are geared toward a centrally managed data environment. By adopting Azure Active Directory permissions groups, the control of access to data will be easier to ensure from source all the way to output reports. Existing reports, when migrated to Power BI, can be consolidated to serve multiple recipients who will only see the data they are permitted access to.

Operational needs for real time data currently put data query loading constraints onto source / operational databases and systems when connections from reports directly to source are required. Moving this to Azure will utilise the specific tooling available for operational and fluid data. Important to note that real-time data when used for reporting is actually slightly delayed, and is in fact near-real-time. For operational reporting this will be evaluated in detail to ensure near-real-time is sufficient. Within the Azure environment there are native data processing, storage and presentation capabilities for such operational data, which are compatible with archive data.

3. A4E and LDDS

Analytics for Everyone (A4E) provides a valuable benchmark from which to guide development of the data transformation project. Already in Azure and using Power BI, A4E utilises some of the best practice methods available. However, the collation of data to a kind of golden dataset that is proposed can build upon the logic that underpins A4E, as a re-engineering exercise rather than a straight lift and shift.

LDDS whilst already in Azure could be linked to for inclusion of historic crime reports. This will need review to establish best-method to architect and implement, to ensure reporting access does not interfere with the search-based drivers in use in the LDDS interface.

4. Proposal

The principal method of achieving reporting success is by de-coupling the report queries from the source databases and systems that supply them. Through a series of transformations that incorporate business logic, the data is curated to best serve the reports. All the while this is segregated from all data sources to ensure there is no additional load on the source systems that would cause slow-downs and record locks. In addition, there would be a reduction in data retrieval that would be significantly reduced when data is readied / cached for reports. In cloud environments the reading from source systems when querying can significantly add to cost. Appropriate architecture for this cost-control purpose is essential.

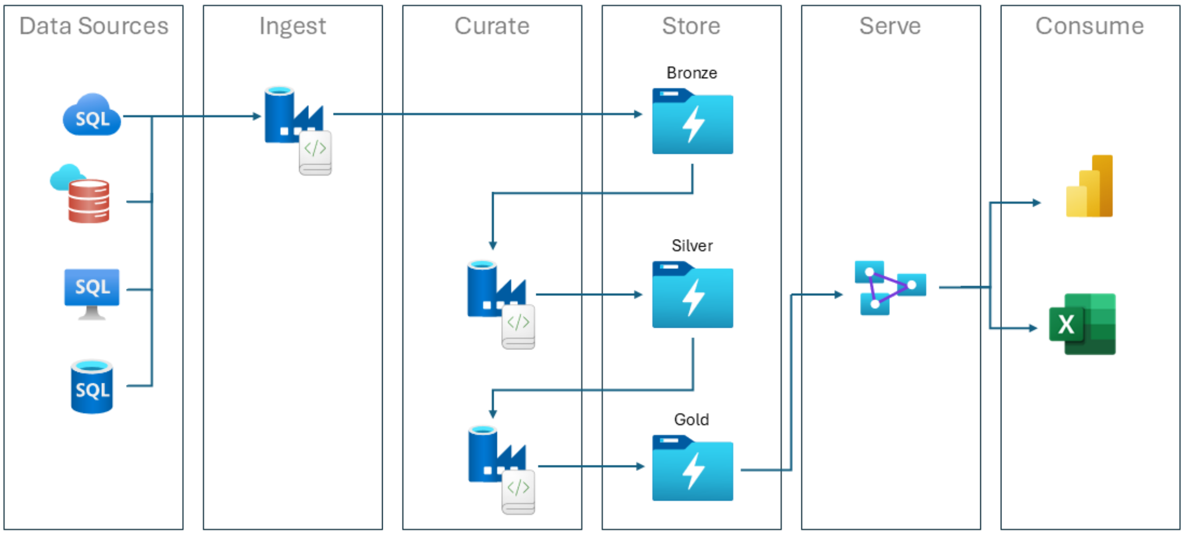
4.1. Data Architecture

Medallion Architecture is the go-to in order to achieve this. It is a simplified structure that uses the data lake to store the data, which is refreshed from the source systems based upon operational timing requirements. The structure is centred around Bronze / Silver / Gold areas (or layers) on the data lake that address each step of the data readiness process for warehousing and reporting.

- Bronze, often referred to as the raw layer, will hold everything ingested from source systems without joining or applying any calculations. These will be updated independently, based upon business needs. The data is ingested as text in order to standardise formatting from different sources that may have different or proprietary data formats.

- The Silver layer is where the calculations are applied to the raw data, to curate the data with logical formatting (date-time, numeric etc) and to only extract meaningful data fields from bronze. This will avoid bringing in a multitude of source systems' ID fields, for example that are only meaningful to the source systems and retain only the data fields necessary for the reporting. This is also where the curated data is 'warehoused', structured for presentation and cached in preparation for presentation in Power BI.

- Gold is where the curated data is assembled, and the warehoused data tables related in readiness for Power BI as Semantic data models. The models that underpin each Power BI report are centrally accessible as directly connected models that can serve multiple reports. This eliminates data model duplication where each report has it's own copy of the same data model as others.



Medallion Architecture

This architectural framework makes use of a data lake to store the data, segregated by source system, to allow for multiple concurrent streams of data ingest. It is similar in principle to having a staging database from which transformations are applied before the data is pushed to a presentation database. These ingest streams are scheduled to run out of office hours (for everything except real-time data) and outside of any system maintenance windows.

Where this method has evolved over staging database use is by storing the data in the medallion layers in Parquet format. Parquet is a self-indexing data format that originates from the world of Big Data. The indexing makes for fast retrieval for querying by Power BI. When the medallion layers are set to utilise parquet files instead of provisioned databases, there are significant cost savings from not having database server infrastructure overheads. The parquet files are essentially the same as database tables, that are stored in a file structure. The medallion layers reference these files in the same way that database tables are queried. This architecture pattern is known as Serverless. Furthermore, as the data being warehoused in the medallion structure is data lake centric, this is known as a Data Lakehouse.

Business Intelligence tools, in our case Power BI, use different processing engines to those that run reporting directly from the source systems and databases. Reliance is placed upon the data being readied as ‘wide’ tables, in the parquet format, comprising of related data that has been prepared before reaching Power BI. The Gold layer holds several specially prepared extracts of the data (from the curated / warehouse data on the silver layer) that are grouped together logically and joined on their relational key fields. These are ready to serve to the users, which they can then consume either as BI reports and dashboards or as data. The data being readied (the related parquet files) is cached by Power BI for the report users. This subsequently speeds up report opening and general use whilst navigating.

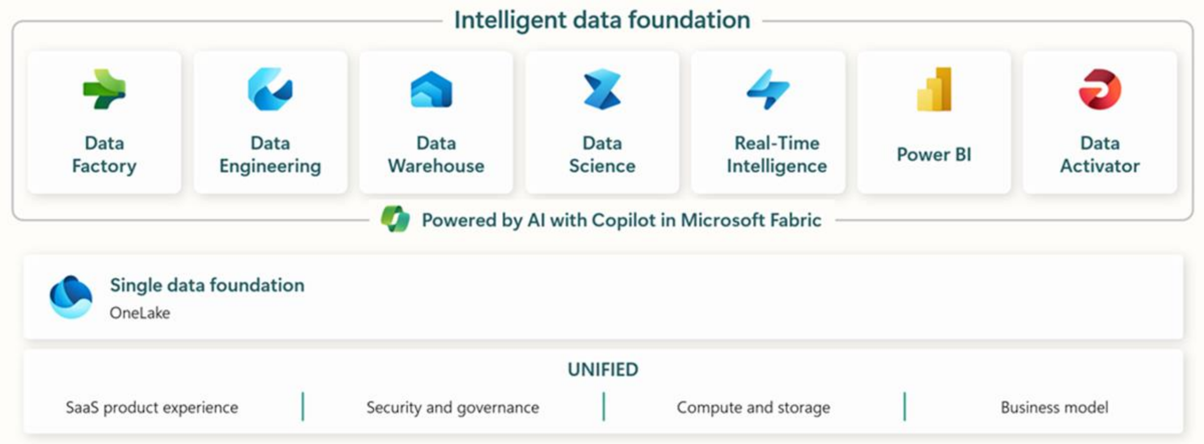
As the data processes move the data through the medallion layers additional metadata from the ingest and curation processes is recorded to enhance auditability. These in turn can be combined with operation logs from Azure and Power BI and are particularly useful when combined with monitoring and governance tools.

Where required, the data that is made available to the end users can be automated to write user updates back into the warehoused data, rather than having the updates exported back to the source system for import or manual entry.

4.2. Fabric

Azure presents a rich and varied set of tools and resources that are suitable for the needs of the data transformation project that include data storage, data processing and presentation. Microsoft Fabric is a source to destination analytics and data platform SaaS solution that brings together the data brought into the medallion architecture’s layers with integrated connectivity to the data tools.

The underpinning feature of Fabric is OneLake, it’s data lake, within which the medallion layers (bronze / silver / gold) are held.



Microsoft Fabric

The components within the Intelligent Data Foundation are essentially resources available within Azure that have native connectivity to OneLake. No need to enable additional resources in Azure such as data factory and databases for warehousing and real-time data provision as they are native components within Fabric.

Data orchestration pipelines are created and executed from the Data Factory, and being integrated in Fabric it no longer requires separate activation in Azure. These will work alongside some script-based processes, leveraging the Big Data technique of using Spark, within Fabric's Data Engineering function. These further allow for some metadata-driven development where Spark notebooks will run some of the repetitive tasks required to ingest and transform data with less code, reduced maintenance and greater scalability than writing code or pipelines for every data source that needs to be ingested and transformed. By determining the data loading and transformation patterns for the data sources and destinations, the metadata framework builds the framework to support each pattern.

Within the toolset there is further extensibility to the data provisioning that manages real-time intelligence. This will allow the use of operational data that can come in live e.g. IoT data, that does not require, nor can it wait for an overnight data refresh. The data can be combined with the warehoused data and can also be served for Power BI in (near) real time separately.

4.3. Data Governance

Microsoft Purview is the data safeguard resource integrated within our Azure proposal. Spanning across the entirety of the end to end of data touchpoints, Purview gives clarity over the data that would otherwise require some additional cataloguing. With the ability to understand the data, wherever it is at any given point from source through to Power BI, the ability is made available to govern and safeguard access to it at each touchpoint, across the data lifecycle.

There are various compliance tools within Purview that range from classifying certain or most of the data to full forensic level auditability.

As a shared feature within our Azure environment Purview acts as a proactive safety net that underpins the entire data lake, tooling and reporting that is architected.

4.4. Security

From an InfoSec viewpoint security and governance underpins Fabric. Azure Active Directory / Entra is the driver behind access permissions that are set from source through to destination. Each touchpoint independently will be controlled by the user roles granted. These are further validateable from Purview's lineage view. Permissions and classifications are searchable and will show only what is permitted to whoever is accessing the data.

For sensitive data that would otherwise be treated as though constrained to a physical standalone machine for isolation, the security equivalent within Fabric's OneLake and AD / Entra is Role Based Access Control (RBAC). If, however the requirement is that some of the data be made secure, as if on a physically segregated machine, then the architecture is adaptable to segregate the data even if that means hosting in a different cloud provider, without compromising the AD / Entra groups and permissions.

Detailed design of the access permissions required across the various data touch points will be part of the design work, that will align with and extend logically what already exists in the current AD groups.

Azure's governance tool, Purview, is the means to apply granular level control over the data within the source to destination journey. It works with the permission structure set in AD / Entra for internal use as well as for externally submitted data.

5. Next steps

The development methodology recommended would begin with a discovery mapping exercise, to record all data systems in use. These in turn will need to be broken down further to their structure where identification of data classification tagging can begin, to permit the governance tools to operate and apply constraints.

5.1. Cloud Adoption Framework

Follow the guidelines set out in Microsoft's framework to align ourselves with recommended actions. Some of these have already been done and may well benefit from acknowledgement within the framework planning tools / task lists. Others may give more granular steerage, for example the documentation that complements the framework beyond this design brief.

As a complementary exercise to the adoption framework a level of data discovery will assist in putting together the framework's documents. Similar articulation of access permissions will be best served by profiling the users and how they move through the business to ensure correct access.

5.2. Mapping document

A detailed list showing all the source systems and the component tables that they contain. This forms the itemisation of the raw data that is ingested to the bronze layer. The transformation processes that include custom calculations will need to be listed here in order that they can be recreated in the data factory and metadata framework.

5.3. User Profiles / Permissions

A matrix of all users who will view the reports and dashboards, showing association to the data that they are permitted to view. These will evolve to include association with business classifications designated in Purview, to align with sensitivity ratings and corresponding restrictions.

This will serve as a component to the security considerations that considers what currently exists, to build upon, as well as the wider InfoSec guardrails.

5.4. Related Documents / Articles

Title	Description
Cloud Adoption Framework CAF Link	Microsoft's lifecycle framework – information and guidance for across the business issues relating to migration.
Well Architected Framework WAF Link	A set of quality-driven tenets and tools to assist in the design and optimisation of workloads.

6. Recommended approach

With guidance from Microsoft's Cloud Adoption Framework and review what has been done already against what still needs to be done, taking into consideration the components and actions that are relevant to Kent and Essex Police.

Some of the components of the frameworks will relate to business analysis that overlap with the overall business case for the data transformation project. These will expand to a more granular level the costs and the associated savings. Others relate to licensing of the products proposed (Fabric) that will form an overlay in the HLD to illustrate the correlation with Azure's extensive cost controls.

6.1. Source to Destination Mapping Data Exercise

The detail component breakdown of source systems will be documented that will allow correct distribution of the systems' tables' data fields, through data design matrices. This will be integral to inclusion within the metadata framework that will allow scripted creation of data factory pipelines to process (ETL) the data between the medallion layers.

6.2. Framework Adoption

Collaboration with business streams to follow the guidance provided by Microsoft's Cloud Adoption Framework, to build upon what has already been done. Standardisation of documents as well as consolidation of architectural diagrams with security, networking, business allocation for budgeting and ongoing cost analysis, and cost benefit analysis alongside baselined known costs.

Adopting cost-saving principles outlined in the Well Architected Framework e.g. script-based data engineering (metadata framework), appropriate provisioning of tooling (reserving instances of Fabric) to give subscription cost efficiencies. Correct sizing and anticipation of operational workloads for data processing that will give more predictability to the natural variations in data engineering costs.

6.3. Discovery Exercise

Recommending the following be set up alongside the data mapping exercise for demonstrable examples of use cases uncovered during discovery:

6.3.1. Multi-Source to Power BI

Connection to Oracle and SQL Server databases, pulling together example / test data to demonstrate the transformation through the medallion architecture to onward presentation in Power BI.

6.3.2. Real-time data

Operational data gathered for KPI reporting of response times to 999 calls, to demonstrate live dashboard and collected data for trend analysis.

6.3.3. Text Analysis

Using Power BI's features for interrogating presented output.

6.3.4. Metadata Framework

Taking data from the data mapping document to create script-based deployment pipelines in the data factory. This will provide insight into the script / coding quality and time saving over manually creating many of the transformation pipelines used in the medallion architecture.

6.3.5. A4E Transition

Will be mirrored for both Kent and Essex data. To show current reporting against new environment.

6.4. High Level Design

Assembly of the next level of detail into a formal design document that will bring together some of the outputs of the discovery exercise, the data mapping, the security grouping required for the new architecture and tooling. Methods of segregating the development and test data environment from the production hardened environment. Continuous integration and deployment protocols using DevOps. This will form the blueprint for development effort.

Tarek Moghul
Data Platform Architecture
IT – Kent & Essex