



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

DATA WAREHOUSE & DATA MINING 01

Muhammad Kamrul Hasan Sohel (2316411)

MODULE CODE: ACCA 6026

STUDENT ID: 2316411

LECTURER NAME: RAJEEV NATH

UNIVERSITY OF WALES TRINITY SAINT DAVID

Abstract

NHS England manages a large and complex healthcare environment involving acute hospitals, community services, mental health services, specialist centres, ambulance pathways, workforce planning and national performance monitoring. Although local Electronic Patient Record systems, Patient Administration Systems, finance systems and workforce platforms support daily transactions, they do not provide a unified and near real-time analytical view across regions, Integrated Care Systems and patient pathways. This report proposes a strategic data warehousing solution that consolidates operational, clinical, workforce, financial and quality data into an integrated warehouse and tailored data marts. The solution is designed to support patient flow analysis, bed utilisation, referral-to-treatment monitoring, workforce capacity planning, financial sustainability and clinical quality improvement. The report also explains OLTP systems, warehouse architecture, ETL, data marts, dimensional modelling, star schema design, data governance and implementation planning using NHS-specific examples and IEEE-style evidence.

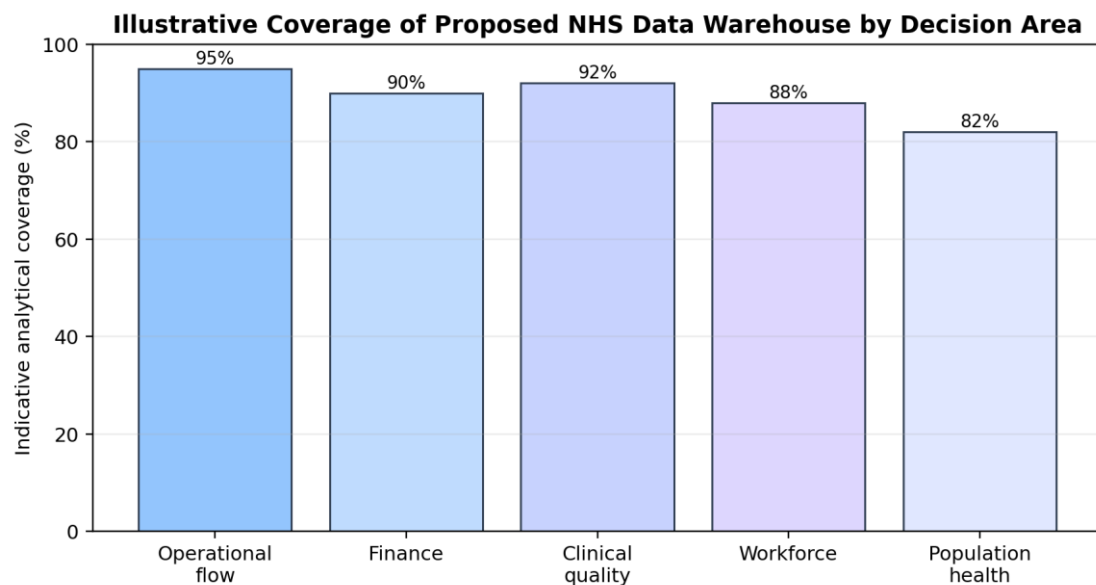
Table of Contents

Abstract.....	0
Introduction.....	3
Interpretation of Analytical Coverage Across NHS Decision Areas:.....	3
Task 1: Data Warehousing Solution for NHS England	4
1.1 OLTP Systems vs Data Warehouse	4
<i>Table 1: Difference between OLTP systems and a data warehouse in NHS England</i>	<i>5</i>
1.2 Data Warehouse Architecture and ETL	5
<i>Figure 2: Proposed NHS England data warehouse architecture integrating internal operational sources</i>	<i>5</i>
<i>Figure 3: ETL process showing how warehouse processing is separated from daily NHS operational systems.....</i>	<i>6</i>
1.3 Data Marts for Management Decision-Making	6
<i>Table 2: Proposed NHS England data marts for management decision-making</i>	<i>7</i>
1.4 OLTP Normalisation vs Data Warehouse Dimensional Modelling.....	7
<i>Figure 4: Star schema for NHS patient activity analytics showing facts and dimensions. [3].....</i>	<i>8</i>
<i>Figure 5: Proposed star schema for the NHS England patient activity data warehouse showing the central fact table and related dimension tables.</i>	<i>Error! Bookmark not defined.</i>
Task 2: Star Schema and Implementation Plan	9
2.1 Source System Identification	9
<i>Table 3: Source systems feeding the NHS England data warehouse</i>	<i>10</i>
2.2 Star Schema Design	10
<i>Figure 6: Multi-star schema design using shared conformed dimensions and aggregate tables. [3]</i>	<i>10</i>
<i>Table 4: Master data, metadata, fact data and aggregate tables in the proposed warehouse</i>	<i>11</i>
2.3 Data Marts and Schema	12
<i>Table 5: Data mart schemas and example KPIs.....</i>	<i>12</i>
2.4 ETL Plan	12
<i>Table 6: Simplified ETL plan for NHS England warehouse tables.....</i>	<i>13</i>
2.5 Data Quality and Governance	13
Conclusion	13
References:.....	14

Introduction

NHS England manages healthcare planning, performance and improvement across a large national system, including teaching hospitals, district general hospitals, community services, ambulance pressures and specialist clinical centres. These services generate large amounts of patient, clinical, operational, workforce, financial and quality data. However, when data remains separated across local systems, senior leaders may struggle to view complete patient pathways or compare performance consistently across regions.

The Department of Health and Social Care highlights that better use of health and care data can improve services if data is managed safely, ethically and transparently [1]. NHS England's Federated Data Platform also shows the importance of connecting operational data to support better care and decision-making [2]. Therefore, this assignment proposes a strategic data warehouse to complement existing NHS systems by creating a secure analytical layer for national, regional and organisational decision-making.



Note: values are design coverage indicators, not actual NHS performance figures.

[3]

Figure 1: Illustrative analytical coverage expected from the proposed NHS warehouse. The values are design indicators only, not actual NHS performance data.

Interpretation of Analytical Coverage Across NHS Decision Areas:

This graph shows the expected analytical coverage of the proposed NHS England data warehouse across five key areas. It shows the strongest support for **operational flow** at 95%, as the warehouse would combine data on admissions, discharges, bed use, waiting lists and

patient movement. **Clinical quality** is also strongly supported at 92%, helping NHS leaders monitor outcomes, readmissions, safety and service performance.

The graph also shows high coverage for **finance** at 90% and **workforce** at 88%, meaning the warehouse would support analysis of costs, budgets, staffing, sickness and agency use. **Population health** has the lowest coverage at 82% because it depends more on external demographic and public health data.

Overall, the graph suggests that the proposed data warehouse would give NHS leaders a wider view of performance across services. However, the figures are only illustrative design indicators and are not actual NHS performance results [4].

Task 1: Data Warehousing Solution for NHS England

1.1 OLTP Systems vs Data Warehouse

An **OLTP system** records and updates daily operational transactions. In NHS England, examples include Electronic Patient Records, Patient Administration Systems, clinical systems, workforce rostering systems and finance platforms. These systems support tasks such as patient registration, admissions, appointments, discharge updates, staff allocation and procedure coding. Their main purpose is to provide fast, accurate and reliable transaction processing [4].

A **data warehouse** has a different role. Inmon describes a data warehouse as subject-oriented, integrated, time-variant and non-volatile, designed to support management decision-making. For NHS England, this means combining historical and cleansed data on patient activity, hospitals, specialties, workforce, finance, bed utilisation and clinical quality. Unlike OLTP systems, the warehouse does not update every live transaction but supports trend analysis, regional comparison and service improvement [5].

The CEO should understand that existing NHS OLTP systems are useful for daily care delivery, but they are not designed for national analytics. For example, a PAS can show whether a patient was admitted or discharged, while a data warehouse can show wider patterns such as discharge delays, RTT waiting time trends or links between staffing costs and bed pressure. Therefore, a warehouse would strengthen national decision-making by integrating operational, clinical and financial data [5] [6].

Aspect	OLTP in NHS England	Data Warehouse in NHS England
Purpose	Record live patient, workforce and financial transactions	Support analysis, reporting and decision-making
Examples	EPR, PAS, rostering, finance and bed systems	Enterprise NHS analytical warehouse and BI layer
Data structure	Highly normalised for accuracy	Dimensional star schemas for reporting
Typical user	Clinicians, administrators and operational staff	CEO, board, finance, medical and regional leaders

Aspect	OLTP in NHS England	Data Warehouse in NHS England
Question answered	Has this patient been discharged?	Which regions have the highest delayed discharge pressure?
Performance risk	Heavy analytics can slow live systems	Analytics run on copied data, reducing disruption

Table 1: Difference between OLTP systems and a data warehouse in NHS England

1.2 Data Warehouse Architecture and ETL

A suitable NHS England data warehouse architecture should use a layered design: source systems, secure extraction, staging, transformation, enterprise warehouse and presentation/data mart layer. This is important because NHS data comes from many systems and regions. A London teaching hospital, a Midlands district general hospital and a rural South West provider may use different local systems, but the national warehouse must transform these different datasets into common structures and definitions.

NHS England Data Warehouse Architecture

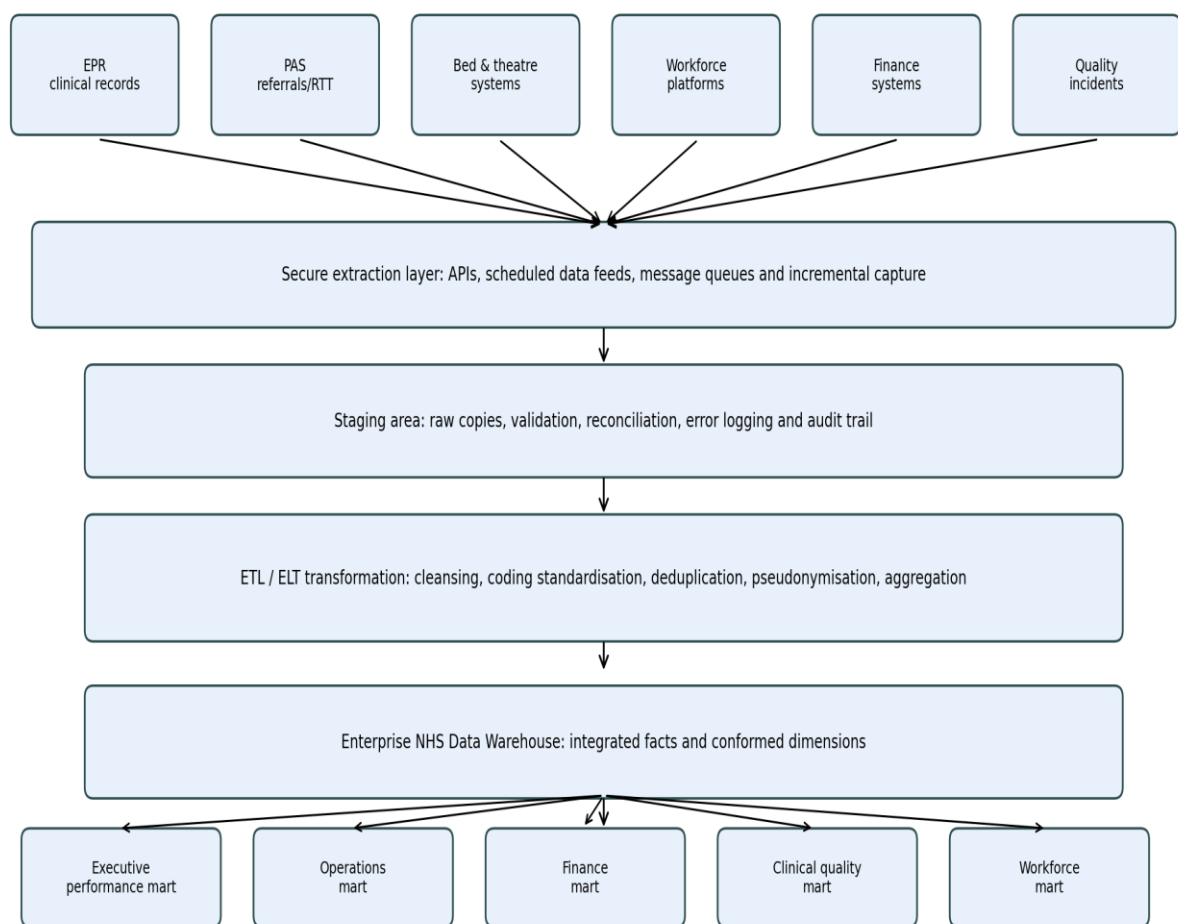


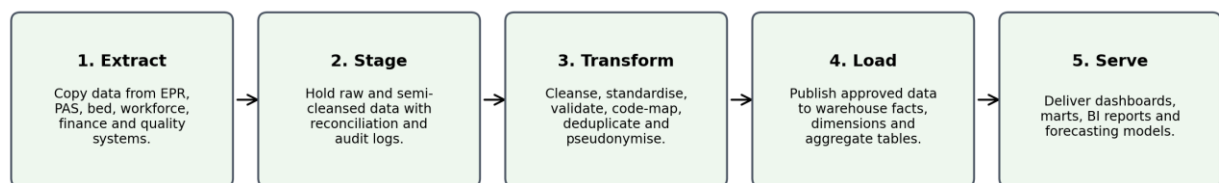
Figure 2: Proposed NHS England data warehouse architecture integrating internal operational sources.

The source layer should collect data from EPR, PAS, theatre, bed management, workforce, finance and quality systems. These sources provide information on clinical care, referrals, admissions, bed occupancy, staffing, costs, incidents and patient experience. External population and deprivation data should also be included to support demand forecasting and health inequality analysis.

ETL means **Extract, Transform and Load**. During extraction, data is copied from source systems through secure APIs, scheduled feeds, database views or message-based interfaces. This protects live EPR and PAS systems because analytical queries are run on controlled warehouse copies rather than operational systems.

During transformation, data is cleaned and standardised. NHS numbers can be validated, duplicate records removed, ICD-10 diagnosis codes mapped, provider and ICS codes standardised, and finance codes aligned with tariff structures. Where direct identification is not needed, patient data should be pseudonymised to reduce privacy risk. Finally, approved data is loaded into fact and dimension tables for dashboards and data.

ETL Process Designed to Avoid Disruption to Daily NHS Operations



Operational systems remain the source of transactions; analytical workloads are moved to the warehouse.

Figure 3: ETL process showing how warehouse processing is separated from daily NHS operational systems.

Healthcare interoperability standards are also important. HL7 FHIR provides a widely used method for representing and exchanging healthcare data, supporting more consistent integration between clinical systems and analytical platforms. NHS England should therefore design the integration layer to accept both structured extracts and standards-based data feeds where possible [3],[9],[10]

1.3 Data Marts for Management Decision-Making

A **data mart** is a smaller subject-specific area taken from the enterprise data warehouse. It helps each management group access the information most relevant to its decisions without using the full national warehouse. For NHS England, key data marts should focus on executive performance, operations, finance, clinical quality and workforce planning.

The **Executive Performance Data Mart** should support the Chief Executive and board by showing high-level indicators such as RTT waiting times, elective backlog, A&E pressure, bed occupancy, ambulance handover delays, cancer performance, staff vacancies, agency use, budget variance and patient safety trends. This would provide a national view by region, ICS, trust and service line, supporting priorities, recovery planning and targeted intervention.

The **Finance Data Mart** should link financial data with clinical and operational activity. It should include budgets, expenditure, tariffs, cost per episode, agency spend, overtime, theatre

utilisation and cost per bed day. This helps explain not only how much is being spent, but also why costs differ between services and regions.

The **Clinical Quality Data Mart** should focus on outcomes and safety, including readmissions, mortality indicators, infections, safety incidents, diagnostic delays, patient satisfaction and compliance with standards. This would help clinical leaders identify quality variation and direct improvement work more effectively.

Data Mart	Primary Users	Core Data	Decision Supported
Executive Performance	CEO and Board	RTT, backlog, beds, workforce, finance and safety indicators	National priorities, escalation and regional intervention
Operations	Regional and operational leaders	Patient flow, bed utilisation, theatre use, discharges and waiting lists	Capacity planning and bottleneck removal
Finance	Finance Director and finance teams	Budgets, tariffs, costs, expenditure, agency spend and cost centres	Resource allocation and cost control
Clinical Quality	Medical Director and clinical governance teams	Outcomes, readmissions, mortality indicators and safety incidents	Quality improvement and patient safety
Workforce	HR and operational leaders	Staffing, sickness, overtime, agency use and vacancies	Workforce planning and agency reduction

Table 2: Proposed NHS England data marts for management decision-making

These data marts also improve information governance because access can be restricted by role. For example, a finance analyst may require cost and activity summaries, while a clinical governance lead may require pseudonymised patient-level quality data. Role-based access is necessary because health data is special category data under UK GDPR [9] [10].

1.4 OLTP Normalisation vs Data Warehouse Dimensional Modelling

OLTP systems normally use normalised database structures. Normalisation separates data into related tables to reduce duplication and maintain data integrity. For example, a PAS may store patient demographics, admissions, consultants, wards, referrals and discharge information in separate linked tables. This design supports accurate updates because one patient record can be amended without repeating the same data across many records.

A data warehouse uses dimensional modelling because the purpose is analysis rather than transaction processing. In a dimensional model, measurable business events are stored in fact tables and descriptive context is stored in dimension tables. This is easier for managers and analysts because the model reflects reporting questions. For example, NHS England may want to analyse length of stay by hospital, specialty, diagnosis, month and age band. A star schema makes this query simpler and faster than joining many normalised OLTP tables.

Star Schema for NHS Patient Activity Analytics

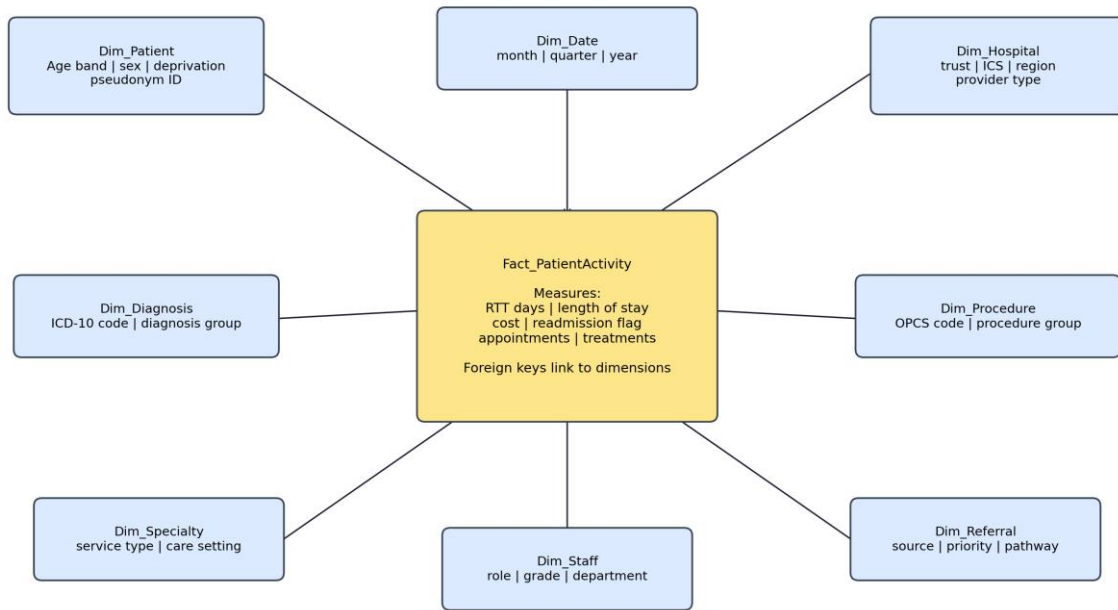


Figure 4: Star schema for NHS patient activity analytics showing facts and dimensions. [3]

In Figure 4, the central **Fact_PatientActivity** table records measurable events such as RTT days, length of stay, cost and readmission flag. Dimension tables describe the context, including patient group, date, hospital, diagnosis, procedure, speciality, staff and referral source. This supports national reporting because the same dimensions can be reused across several fact tables. Therefore, OLTP normalisation is best for reliable daily care transactions, while dimensional modelling is best for national performance analytics [9] [11].

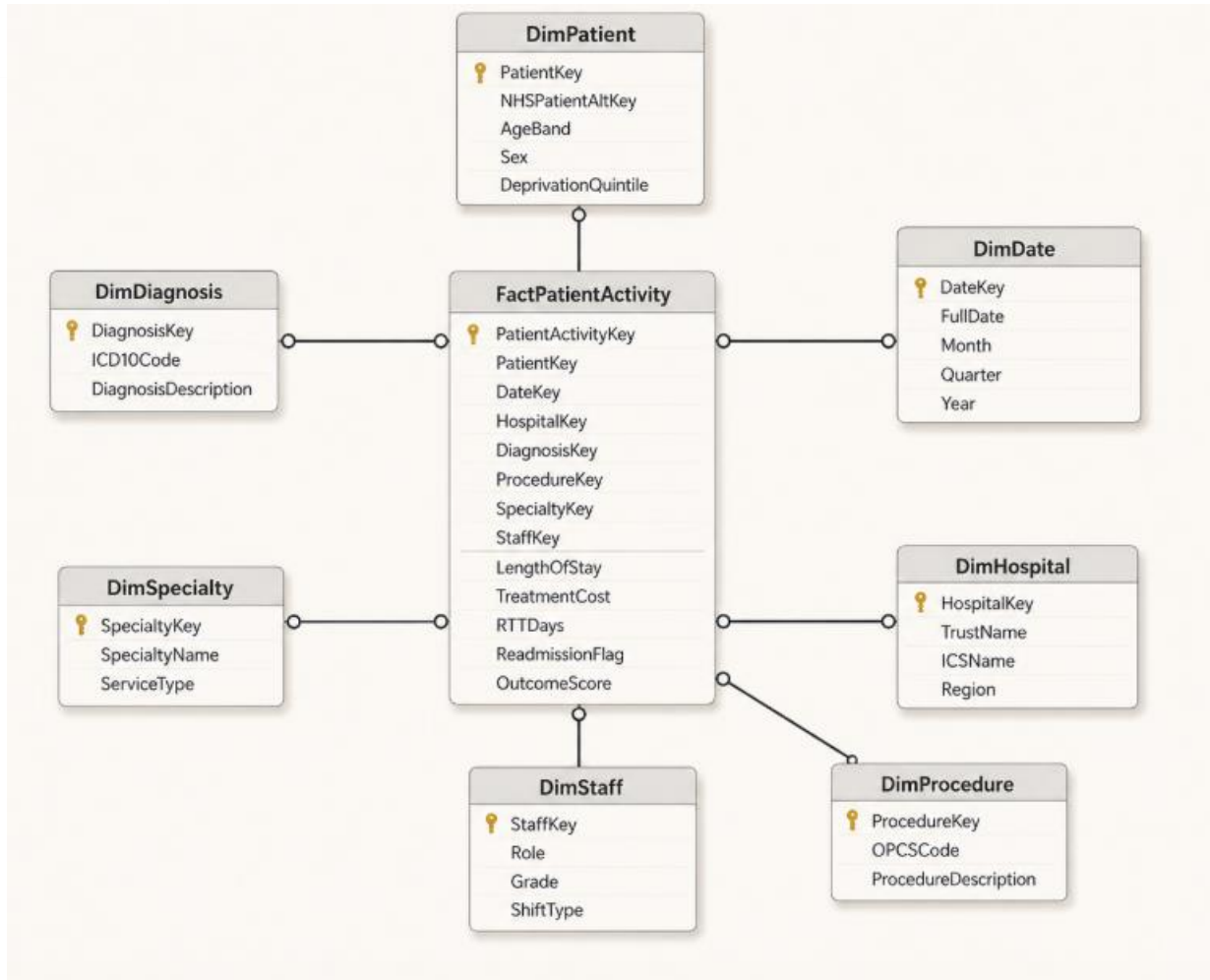


Figure 5: Proposed star schema for the NHS England patient activity data warehouse showing the central fact table and related dimension tables.

[3]

This star schema shows how NHS patient activity data can be structured for analysis. The central **FactPatientActivity** table stores key measures such as length of stay, treatment cost, RTT days, readmission status and outcome score. It connects to dimension tables including **DimPatient**, **DimDate**, **DimHospital**, **DimDiagnosis**, **DimProcedure**, **DimSpecialty** and **DimStaff**. This design supports faster reporting and helps NHS decision-makers analyse patient care, costs, hospital performance and clinical outcomes.

Task 2: Star Schema and Implementation Plan

2.1 Source System Identification

The warehouse should receive internal and external data. Internal sources include EPR systems for clinical records, PAS systems for referrals and RTT pathways, clinical systems for diagnostics and treatments, theatre systems for session utilisation, bed management systems for ward occupancy, workforce systems for staffing and absence, finance systems for tariffs and expenditure, and incident systems for patient safety. External sources include ONS population data, deprivation data, national tariff data, published NHS statistics and public health information. These sources are needed because operational performance is affected by both internal capacity and external demand [12].

Source System	Example Data	Warehouse Use
EPR	Diagnoses, outcomes, clinical observations	Clinical quality and outcome analysis
PAS	Referrals, appointments, admissions, discharges and RTT	Waiting list, patient flow and length of stay analysis
Theatre systems	Sessions, cancellations and utilisation	Elective recovery and capacity planning
Bed systems	Occupied beds, available beds and delayed discharges	Bed utilisation and patient flow management
Workforce systems	Shifts, sickness, overtime and agency use	Workforce capacity planning
Finance systems	Budgets, tariffs, expenditure and cost centres	Financial sustainability and cost analysis
Incident systems	Safety incidents, complaints and patient experience	Quality governance
External data	Population, deprivation and public health data	Demand forecasting and health inequality analysis

Table 3: Source systems feeding the NHS England data warehouse

2.2 Star Schema Design

The recommended design is a multi-star warehouse rather than one single large table. NHS England has several major analytical subjects: patient activity, bed utilisation, workforce capacity, finance and clinical quality. Each subject should have a fact table supported by shared conformed dimensions. This design follows dimensional modelling principles and supports consistent analysis across departments.

Multi-Star Warehouse Design for NHS England

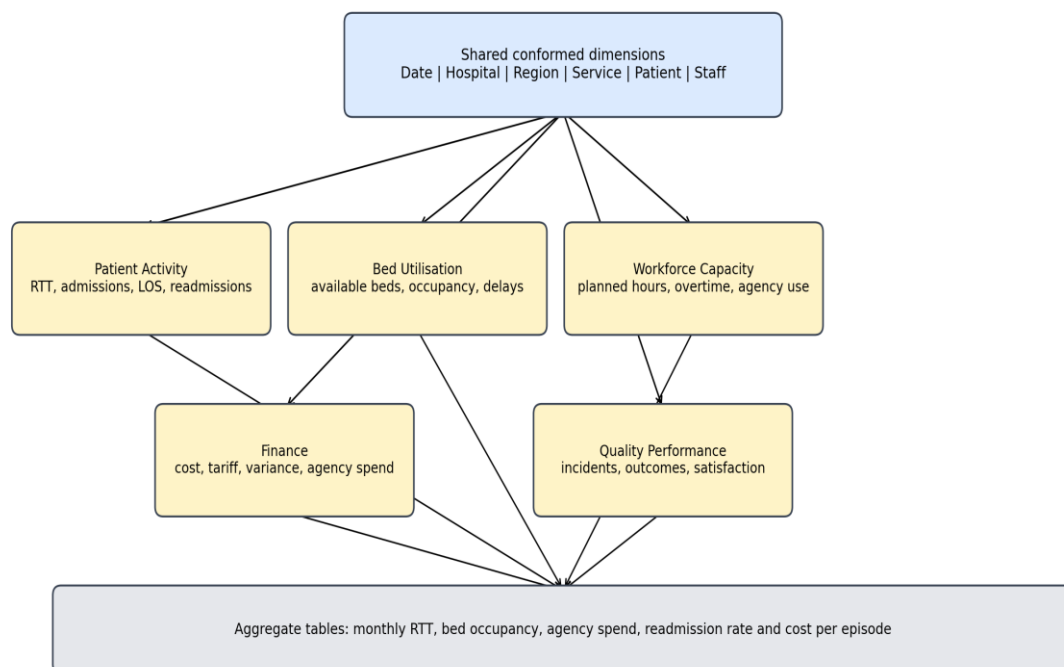


Figure 6: Multi-star schema design using shared conformed dimensions and aggregate tables. [3]

The first fact table should be **Fact_PatientActivity**, which records referrals, admissions, outpatient appointments, procedures, discharges and follow-up activity. Key measures include RTT days, length of stay, appointment count, procedure count, discharge delays, treatment cost and readmission flag. By linking these measures to dimensions such as Patient, Date, Hospital, Region, Diagnosis, Procedure, Specialty and Referral Source, NHS England can analyse elective backlog, emergency pressure, outpatient demand and readmission patterns.

The second fact table should be **Fact_BedUtilisation**, which records daily bed availability and occupancy across wards, hospitals, service types and regions. Measures such as available beds, occupied beds, escalation beds, delayed discharges, occupancy percentage and average length of stay would help identify patient flow bottlenecks across different NHS settings.

The third fact table should be **Fact_WorkforceCapacity**, covering planned hours, worked hours, overtime, agency hours, sickness, vacancies and staff-to-patient ratios. This would allow NHS England to assess whether staffing pressures are linked to waiting times, bed pressure or patient safety outcomes.

The fourth fact table should be **Fact_Finance**, which records treatment costs, tariff income, budgets, expenditure, variance, agency spend, cost per episode and cost per bed day. This supports financial analysis by connecting cost differences to clinical activity, staffing and operational pressures.

The fifth fact table should be **Fact_QualityPerformance**, covering readmissions, mortality indicators, safety incidents, infection indicators, complaints, patient satisfaction and target compliance. This would help clinical leaders identify quality variation and target improvement resources more effectively.

Table Type	Example Tables	Purpose
Master data	Dim_Patient, Dim_Staff, Dim_Hospital, Dim_Department, Dim_ServiceType	Provides consistent reference entities used across facts
Metadata	Data_Dictionary, Source_System, Refresh_Log, Data_Quality_Rule	Explains definitions, lineage, ownership and refresh frequency
Fact data	Fact_PatientActivity, Fact_BedUtilisation, Fact_WorkforceCapacity, Fact_Finance, Fact_QualityPerformance	Stores measurable events and operational performance
Aggregate data	Agg_Monthly_RTT, Agg_Bed_Occupancy, Agg_Agency_Spend, Agg_Readmission_Rate	Speeds board reporting and monthly performance dashboards

Table 4: Master data, metadata, fact data and aggregate tables in the proposed warehouse

The use of conformed dimensions is particularly important. If Dim_Hospital, Dim_Date, Dim_Region and Dim_ServiceType are shared across all facts, NHS England can compare bed occupancy, workforce shortages, cost and quality outcomes for the same organisation and month. This creates a stronger single version of truth and reduces inconsistent reporting. [4], [13]

2.3 Data Marts and Schema

The Operations Data Mart should combine patient activity and bed utilisation data to monitor waiting lists, RTT days, emergency admissions, discharge delays, occupancy, theatre use and cancellations. This supports operational decisions on patient flow, escalation and backlog recovery.

The Finance Data Mart should link finance, patient activity and workforce data, including treatment costs, tariff income, budget variance, agency spend, overtime and cost per bed day. This helps finance teams understand the reasons behind overspending.

The Clinical Quality Data Mart should combine quality and patient activity data, including readmissions, mortality indicators, safety incidents, infections, patient satisfaction and outcomes. This supports clinical governance and quality improvement.

The Workforce Data Mart should include staffing, roles, shifts, sickness, overtime, vacancies and agency use. The Executive Data Mart should summarise all marts into board-level dashboards using red, amber and green indicators. [4], [13]

Data Mart	Main Schema Content	Example KPI
Operations Mart	Patient activity facts, bed facts, ward, hospital, region and service dimensions	RTT compliance, bed occupancy and delayed discharge count
Finance Mart	Finance facts, cost centres, procedure, hospital and workforce dimensions	Cost per patient episode and budget variance
Clinical Quality Mart	Quality facts, diagnosis, procedure, incident type and hospital dimensions	Readmission rate and patient safety incident trend
Workforce Mart	Workforce facts, staff, role, department and hospital dimensions	Agency hours, sickness absence and vacancy rate
Executive Mart	Monthly aggregate indicators from all subject areas	National RAG dashboard and regional comparison

Table 5: Data mart schemas and example KPIs

2.4 ETL Plan

The ETL plan should use **incremental extraction**, so only new or updated records are copied from operational systems. Data should first move into a **staging area**, where checks for validation, completeness and reconciliation are completed.

The transformation stage should include data cleansing, standardisation, deduplication, pseudonymisation, coding alignment and aggregation. Once validated, data should be loaded into **dimension tables first** and then into **fact tables**. This approach protects live NHS systems, improves data quality and creates a clear audit trail for governance [14].

Target Table	Source	Key Transformations
Dim_Patient	EPR/PAS	Validate NHS number, deduplicate, assign pseudonym key and group age bands
Dim_Hospital	NHS organisation reference data	Standardise trust, ICS, provider type and region
Dim_Diagnosis	Clinical coding systems	Map to ICD-10 and diagnosis groups
Fact_PatientActivity	PAS/EPR	Calculate RTT days, length of stay, readmission flag and discharge delay
Fact_BedUtilisation	Bed management systems	Aggregate occupied and available beds by day, ward and hospital
Fact_WorkforceCapacity	Rostering/HR systems	Standardise roles and calculate overtime, sickness and agency hours
Fact_Finance	Finance systems	Map cost centres, tariffs, budgets and expenditure
Fact_QualityPerformance	Incident and outcome systems	Standardise incident type, outcome measure and reporting period

Table 6: Simplified ETL plan for NHS England warehouse tables

2.5 Data Quality and Governance

Data quality and governance are critical because warehouse outputs may influence patient care, funding and public accountability. NHS England should define ownership, data standards, validation rules, refresh schedules, access permissions and audit procedures. Health data is special category data under UK GDPR, so privacy, role-based access, encryption, pseudonymisation and audit logging are essential. A data dictionary should define indicators such as RTT, bed occupancy, readmission and cost per episode to prevent inconsistent reporting. Data quality dashboards should monitor completeness, accuracy, consistency, timeliness and validity during ETL and ongoing operation [15].

Conclusion

NHS England requires a strategic data warehouse to integrate patient, clinical, operational, workforce, financial and quality data into one secure analytical environment. Unlike OLTP systems, the warehouse would support national reporting, trend analysis and evidence-based decision-making without disrupting daily EPR and PAS operations. The proposed design uses layered architecture, ETL processing, star schemas, conformed dimensions, aggregate tables and data marts for executive, operational, financial, clinical and workforce needs. With strong governance, privacy controls and data quality management, the warehouse can provide a trusted single version of truth and support better planning, efficiency, cost control and patient outcomes.

References:

- [1] Department of Health and Social Care, “Data saves lives: reshaping health and social care with data,” GOV.UK, Jun. 2022.
- [2] NHS England, “NHS Federated Data Platform,” NHS England, 2026.
- [3] JGraph Ltd., “diagrams.net,” *diagrams.net*. [Online]. Available: <https://app.diagrams.net/>. Accessed: Jul. 22, 2026.
- [4] NHS England, “NHS England public data gateway,” *NHS England*. [Online]. Available: <https://data.england.nhs.uk/>. Accessed: Jul. 25, 2026.
- [5] J. Serra, *Deciphering Data Architectures: Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse, and Data Mesh*. Sebastopol, CA, USA: O’Reilly Media, 2024.
- [6] T. Knezevic Ivanovski, S. Honap, R. Matic, S. Markovic, and L. Peyrin-Biroulet, “Building a healthcare data warehouse: considerations, opportunities, and challenges,” *Frontiers in Digital Health*, vol. 7, Art. no. 1691142, 2025, doi: 10.3389/fdgh.2025.1691142.
- [7] Department of Health and Social Care, “Data saves lives: reshaping health and social care with data,” GOV.UK, Jun. 13, 2022. [Online]. Available: <https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data>. Accessed: Jul. 25, 2026
- [8] G. Gavrilov, E. Vlahu-Gjorgievska, and V. Trajkovik, “Healthcare data warehouse system supporting cross-border interoperability,” *Health Informatics Journal*, vol. 26, no. 2, pp. 1321–1332, 2020, doi: 10.1177/1460458219876793.
- [9] J. Serra, *Deciphering Data Architectures: Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse, and Data Mesh*. Sebastopol, CA, USA: O’Reilly Media, 2024.
- [10] Information Commissioner’s Office, “What is special category data?,” ICO. [Online]. Available: <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/special-category-data/what-is-special-category-data/>. Accessed: Jul. 26, 2026.
- [11] A. Kumar, A. Mishra, and S. Kumar, *Architecting a Modern Data Warehouse for Large Enterprises: Build Multi-cloud Modern Distributed Data Warehouses with Azure and AWS*. Berkeley, CA, USA: Apress, 2024.
- [12] NHS England, “Referral to Treatment (RTT) waiting times,” *NHS England*. [Online]. Available: <https://www.england.nhs.uk/statistics/statistical-work-areas/rtt-waiting-times/>. Accessed: Jul. 31, 2026.
- [13] S. Lyu, S. Craig, and G. O’Reilly, “The development and use of data warehousing in clinical settings: a scoping review,” *Frontiers in Digital Health*, vol. 7, Art. no. 1599514, 2025, doi: 10.3389/fdgh.2025.1599514.
- [14] G. Gavrilov, E. Vlahu-Gjorgievska, and V. Trajkovik, “Healthcare data warehouse system supporting cross-border interoperability,” *Health Informatics Journal*, vol. 26, no. 2, pp. 1321–1332, 2020, doi: 10.1177/1460458219876793.

[15] Information Commissioner's Office, "What is special category data?," ICO. [Online]. Available: <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/special-category-data/what-is-special-category-data/>. Accessed: Jul. 21, 2026.

Appendix:

