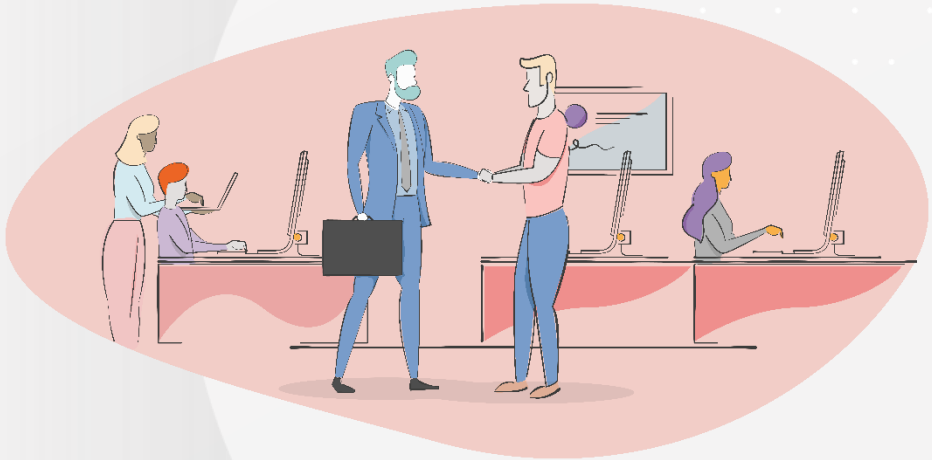
Database administration and optimisation



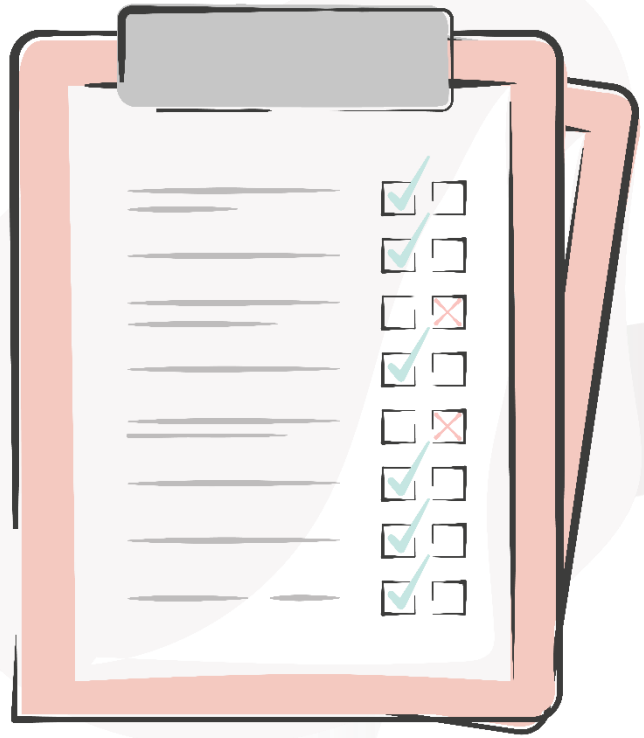
L5 Data Engineer Higher Apprenticeship
Module 2 / 12 (“**Databases and Data Lakes**”)
Topic 4 / 5

rev. 1 (2024)

Webinar agenda

This webinar will cover the following:

- Data profiling
- Database archiving
- Query profiling
- Query optimisation
- Recovery
- Indexing
- Security
- Outages

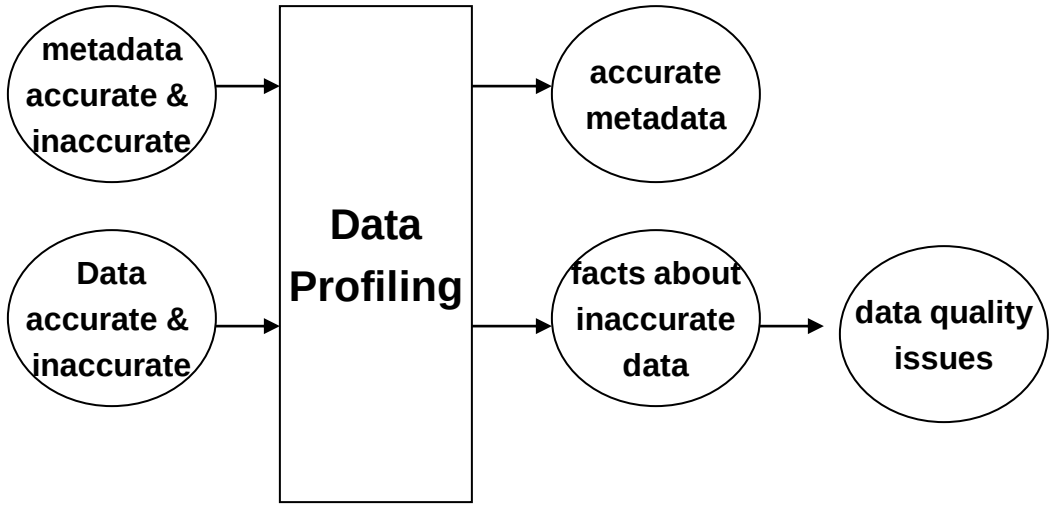


Data Profiling

BPP

Definition

Data Profiling is the application of data analysis techniques to existing data for the purpose of determining the actual content, structure, and quality of the data.



Role of Metadata and Data Rules

- Must define what constitutes correct
- Traditional metadata is only part of definition
- Traditional metadata is often inaccurate and/or incomplete
- Data rules exist whether known or not
- Data rules exist whether enforced or not
- Profiling can validate metadata and known data rules
- Can be used to correct or enhance metadata and data rules
- Can be used to discover additional data rules
- Can test adherence to data rules

BPP

Where is Data Profiling Used?

Database Quality Improvement Program

- Traditional Six Sigma Like Program
- Recursive application of data quality assessment
- Based on historical success of companies who have used it

Support Consolidation of Databases after mergers and acquisitions

- Dramatically reduce cost and time to complete projects
- Improve quality of data in resulting system

Support data integration functions for data warehousing/ business intelligence data stores

- Develop processes to cleanse data in transit
- Improve quality of data in information intelligence stores

BPP

Database Quality Program

Data Quality Discovery

- Data Profiling
- Incident Investigation

Data Quality Correction

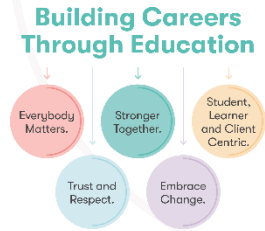
- Name & Address Cleansing De-duplication
- Re-verification
- Data Quality Issue Formulation & Tracking

Data Quality Prevention

- Data Quality Monitoring
- Business Process Improvements
- Application Software Improvements

BPP

Data Profiling Functions



Column Examination

- List all values in column with frequency of occurrence
- Show high and low values
- Determine true data type
- Determine degree of uniqueness
- Determine encoding patterns used, frequency of each pattern
- Compute values: AVG, SUM, MEDIAN, STD DEVIATION

Row Examination

- Find all primary key candidates (single or multi-column)
- Find intra-row column dependencies (find denormalization instances)
- Find multi-column value relationships
  - Value ordering rules
  - NULL value dependencies

Multi-table Examination

- Find matching columns across tables
- Match by column name, data type
- Match by values
- Find primary/foreign key pairs (single and multi-column)
- Determine 1-1, 1-M, 1-0, M-1, M-M, 0-1 rules
- Find primary values not found in secondary tables

Test User provided data rules

BPP

What Types of Problems can be identified for values in a column?

Invalid Values

- Missing values when should not be missing
- Values out of range or not in domain of expected values
- Value in one column not possible when combined with values in one or more other columns
- Obviously wrong when looked at
    - Name = Donald Duck
    - Address = 1600 Pennsylvania Avenue

Valid Values

- Distribution of values unexpected
    - too many of one or more values
    - too few of one or more values
- Value is incompatible with other values in other columns
- Multiple values mean same thing

BPP

What Types of Problems can be identified for values in a column?

Synonyms (multiple representations for same value)

Multiple ways to express same value

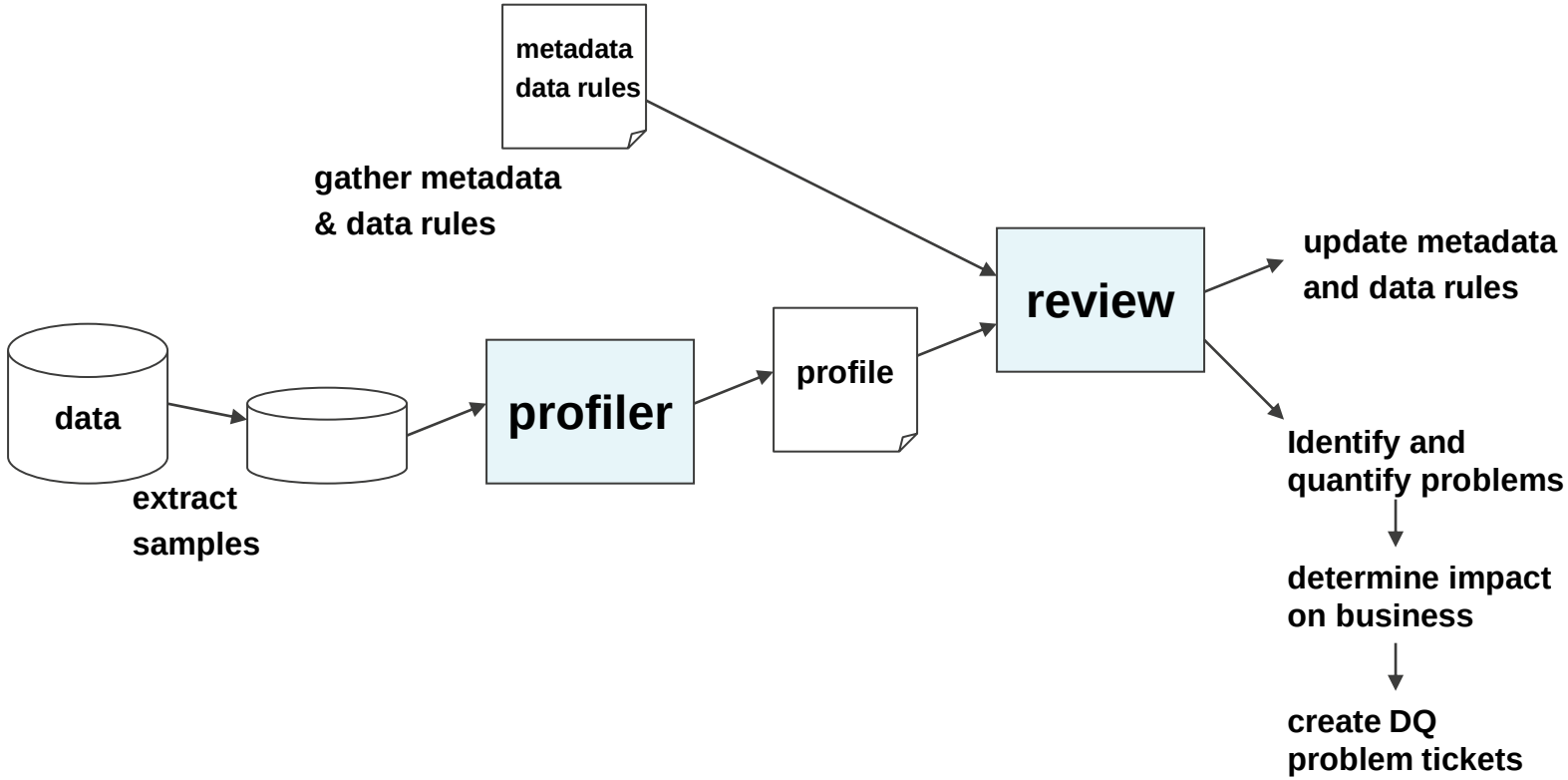
- Date formats
- Number formats
- Use of case on character values

Cases where the value of a column can be determined from the values in one or more other columns

- Explicitly: through a rule
- Correlation against known combinations

BPP

Data Profiling Solution Architectures



Data Profiling Functions

Home Grown

- Inefficient
- Directed SQL; not generic
- Easy Algorithms
- Few Surprises uncovered

Vendor Tools

- More sophisticated functions
  - third normal form discovery
  - foreign key discovery
  - multiple pattern recognition
- More management of results

BPP

Business Case Basics

Dollar savings

- Fewer Operational Mistakes
  - Reduced rework
  - Reduced customer returns
- Improved Analytic Results
  - Better decision making from Analytics

Improved Operations

- Fewer operational glitches due to wrong data values
- Reduced time to complete projects

Risk Reduction

- Reduced exposure to catastrophic events
- Reduced exposure to legal scrutiny of data

Data Quality problems can cost a company 15-25% of bottom line profit.

It’s all indirect: you cannot determine what value you will get until after you have profiled data, studied results and used it to some advantage.

How to get started

- **Start with a pilot data quality assessment of a critical data resource**
  - Small
  - Targeted: high value consequences of poor data
  - “see what we find”
- **Possibly start with a renovation/ consolidation/ integration project**
  - Process to establish source level metadata and understand data content
- **Need to Quantify Results**
- **Use Success to Expand Use of Technology**

BPP

Why is it Neglected?

- **Problems are not on Top Ten List of IT Concerns**
- **Cannot predict value: cannot build business case in advance**
  - Don’t know value of a data quality improvements until after you discover the problems
  - Value is hard to quantify: does not appear until quality problems acted upon
- **Perception that you don’t need it**
- **Exposes corporate/ IT weaknesses**
- **No Clear Owner of Data Quality**
- **Lack of education on value, cost to implement**
  - Management education
  - Technical Staff education

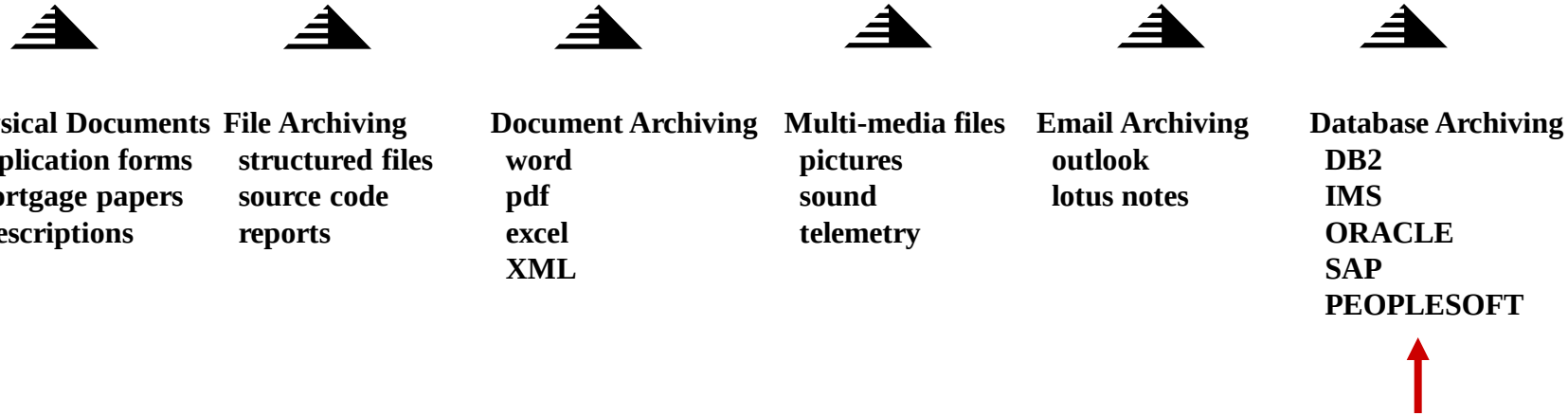
BPP

Database Archiving

BPP

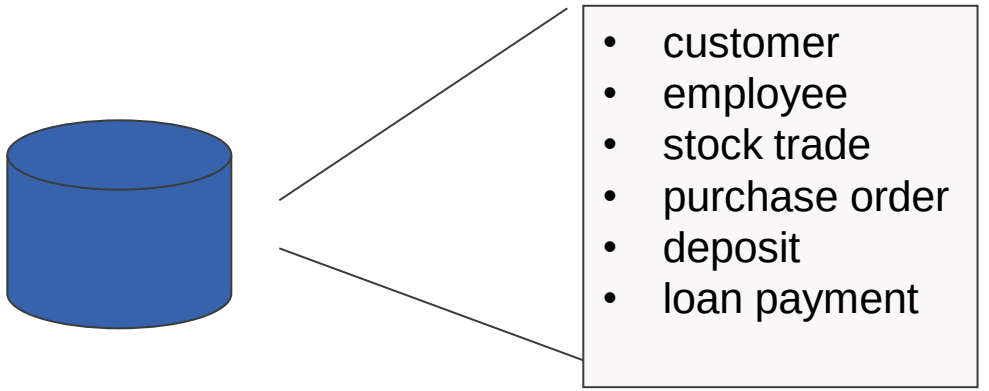
Definition

The process of removing selected data items from operational databases that are not expected to be referenced again and storing them in an archive database where they can be retrieved if needed.



Business Records: the Archive Unit

- You don’t archive databases; you archive data from databases.
- A Business Record is the data captured and maintained for a single business event or to describe a single real world object.
- Databases are collections of Business Records.
- Database Archiving is Records Retention.



Data Retention

The requirement to keep data for a business object for a specified period of time. The object cannot be destroyed until after the time for all such requirements applicable to it has past.

Business Requirements

Regulatory Requirements

The Data Retention requirement is the longest of all requirement lines.

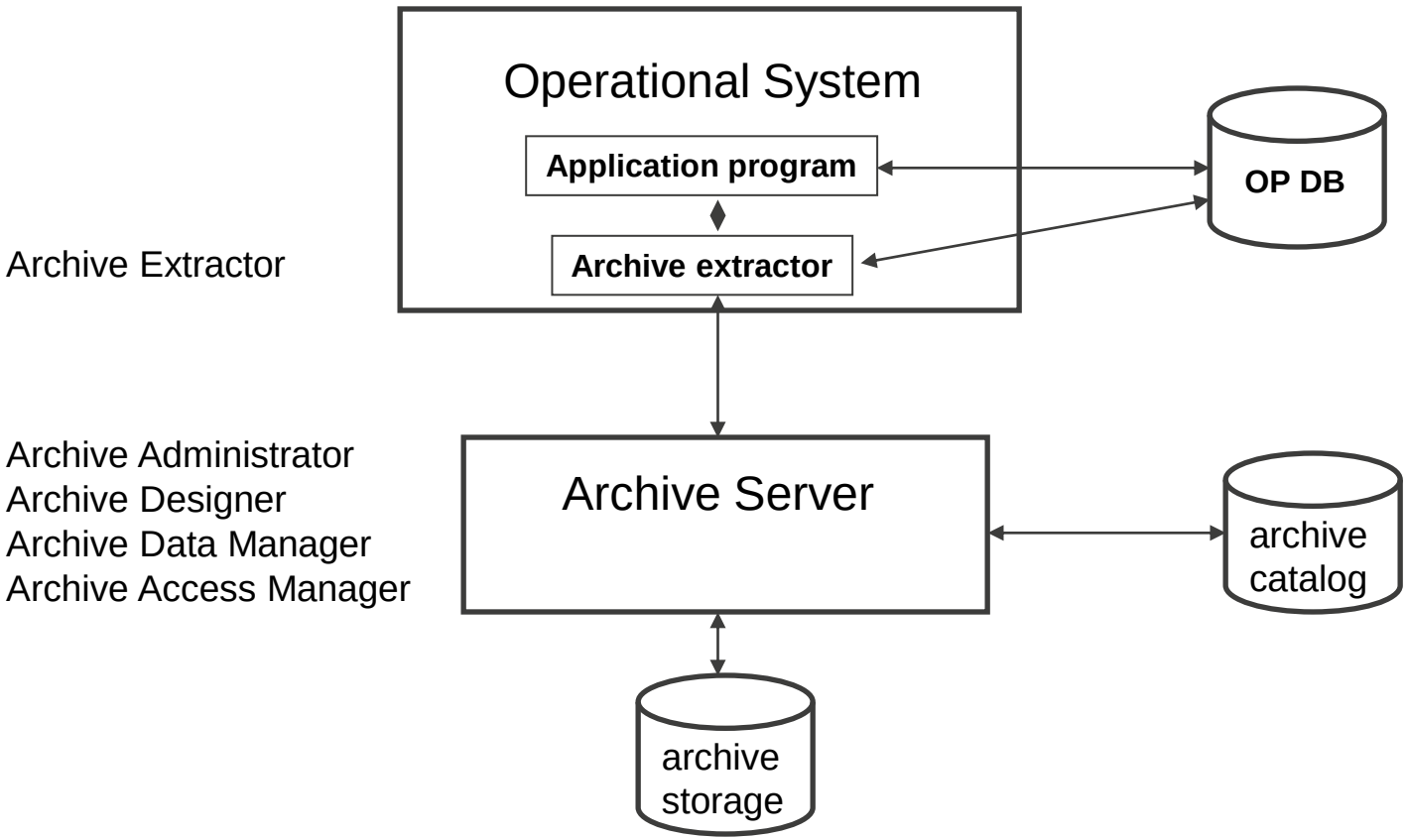
BPP

Data Retention

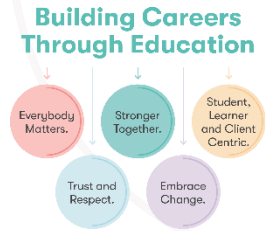
- Retention requirements vary by business object type
- Retention requirements from regulations are exceeding business requirements
- Retention requirements will vary by country
- Retention requirements imply the obligation to maintain the authenticity of the data throughout the retention period
- Retention requirements imply the requirement to faithfully render the data on demand in a common business form understandable to the requestor
- The most important business objects tend to have the longest retention periods
- The data with the longest retention periods tends to be accumulate the largest number of instances
- Retention requirements often exceed 10 years.
- Requirements exist for 25, 50, 70 and more years for some applications

BPP

Architecture of Database Archiving



Archive Staff



Database Archive Specialist

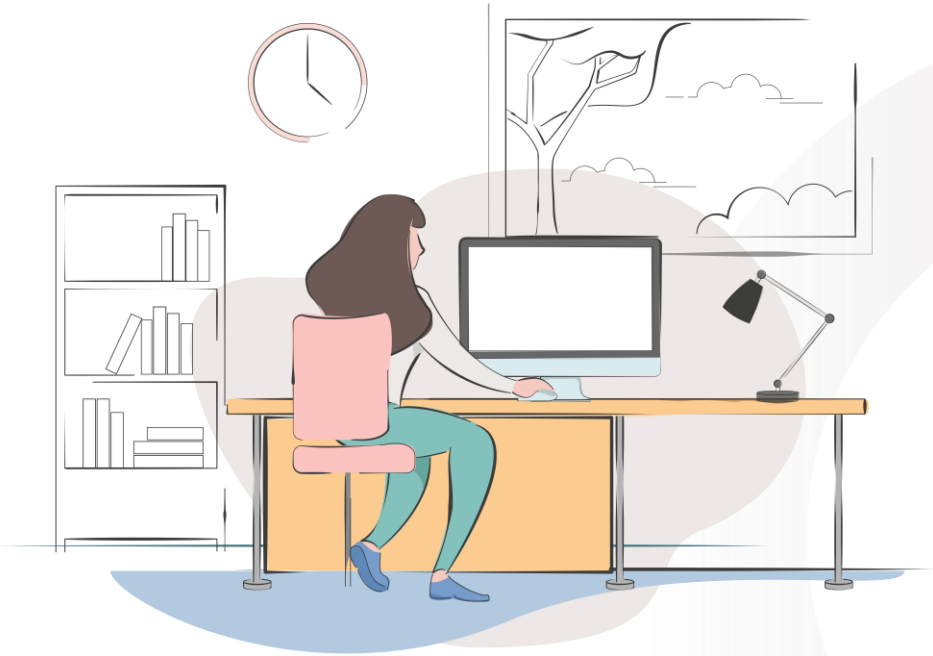
- Received education on database archive design and implementation
- Knows tools available
- Experienced
- Full time job

Database Archive Administrator

- Received education on database archiving administration
- Full time job

Supporting Roles

- Storage Administrators
- Database Administrators
- Data Stewards
- Security Administrators
- Compliance staff
- IT management
- Business Unit Management
- Legal
- Records Management



BPP

Archive Designer Component

Metadata

- Capture current metadata
- Validate it
- Enhance it
- Design archive storage format

Data

- Define business records to be archived
- Define source of data
- Define data structures within operational system
- Define reference data needed to include with it
- Define archive format of data

Policies

- Define extract policy (when a record becomes inactive)
- Define operational disposal policy (when to remove from operational database)
- Define storage policy (how to protect data in archive)
- Define discard policy (when to remove from archive)

BPP

Archive Designer Component

Extractor process

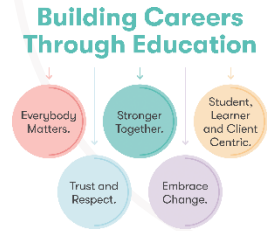
- Verify consistency with design metadata
- Extract data as defined in designer
- Mark or delete from operational database as defined in designer
- Pass data to archive data manager
- Keep audit records on everything done
- Do not impact operational performance
- Support interruptions with transaction level recovery
- Support restart
- Finish scans within acceptable time periods

Scheduling

- Establish periodic executions
- Find non-disruptive periods
- Be consistent

BPP

More reasons for Archiving



All data in operational db

most expensive system
most expensive storage
most expensive software

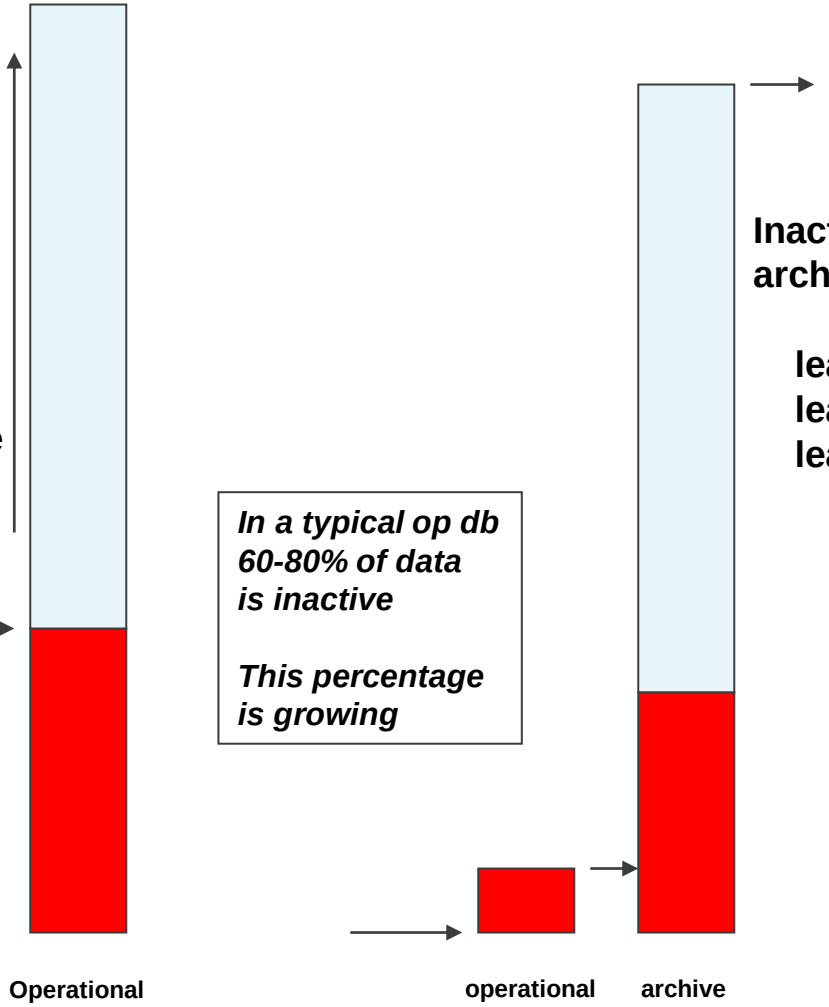
Size Today

In a typical op db 60-80% of data is inactive

This percentage is growing

Inactive data in archive db

least expensive system
least expensive storage
least expensive software



BPP

More reasons for Archiving

- **Look for and compute difference in storage costs**
  - front-line vs archive storage
  - byte counts differences between operational and archive

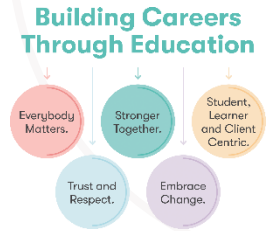
- **Look for and compute difference in system costs**
  operational vs archive systems
  are operational system upgrades avoided
  are software upgrades avoided
  can systems be eliminated for application
  can software be eliminated for application

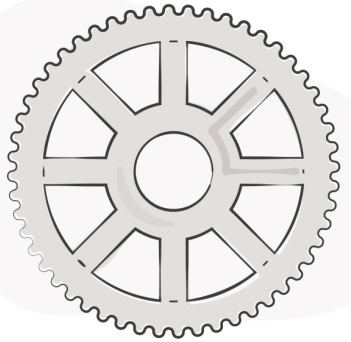
- **Look for savings on people costs**
  - can people be eliminated or redirected for retired applications

- **Potential savings on changes/ application renovations**
  - simplification of design
  - elimination of data conversions

BPP

Operational Efficiency Impacts

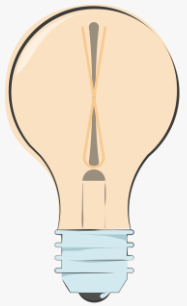




- **Will operational performance be enhanced with less data**
- **Will utility time periods be reduced (backup, reorganization)**
  - fewer occurrences needed
  - less data to process each time
- **Will recovery times be reduced and what is that worth**
  - interruption recoveries
  - disaster recoveries
- **Will implementation of data structure changes be improved**
  - avoided
  - reduced amount of data to unload/modify/reload

BPP

Business case summary



- Database Archiving solutions generally provide for lower cost software, can use lower cost storage more efficiently, and run on smaller machines.

- Each business case is different
  - Many factors can be used in building business case
  - Seen an application justified on storage costs alone
  - Seen an application justified on disaster recovery time alone
  - Seen an application justified on better data security alone

- Each organisation will have many potential applications

- Having a database archiving practice can create synergies across many applications thus adding more value

BPP

SQL Query Optimisation

- NEVER: `SELECT * FROM mytable`
- Learn about query profiling
- Learn about recovery

BPP

SQL Query Optimisation

Query profiling involves collecting detailed information about query execution.

Helps identify slow-running queries.

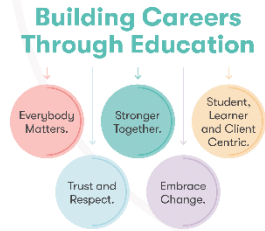
Reveals inefficiencies in query logic and database design.

Tools for Query Profiling:

- **MySQL**: **EXPLAIN**, **SHOW PROFILE**
- **PostgreSQL**: **EXPLAIN ANALYZE**
- **SQL Server**: Query Execution Plans

BPP

Query profiling step-by-step



Step 1

Collecting Execution Data:

- Use profiling tools to gather data on query execution plans and resource usage.
- Example (MySQL): EXPLAIN SELECT * FROM employees WHERE department_id = 10;

Step 2

Analyzing Execution Plans

- Review the query execution plan to understand how the database processes the query.
- Look for signs of inefficiency, such as full table scans or missing indexes.

Step 3

Monitoring Resource Usage

- Track CPU, memory, and I/O usage during query execution.
- Identify queries that consume excessive resources.

Step 4

Identifying Bottlenecks

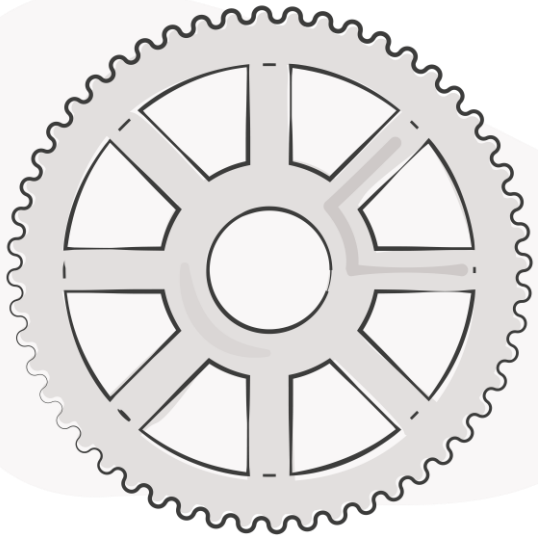
- Pinpoint stages in the query execution where delays occur.
- Focus on optimizing these critical areas.

BPP

Query Optimisation

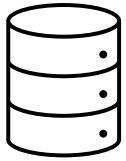
- Query optimization involves applying techniques to improve query performance based on the insights gained from profiling.

- Optimizing queries ensures faster data retrieval and more efficient use of database resources.



Optimisation: Indexing

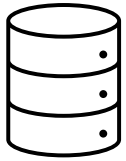
- Create indexes on columns frequently used in WHERE clauses and joins.
- Avoid over-indexing, which can slow down write operations.
- Example: CREATE INDEX idx_department_id ON employees(department_id);



BPP

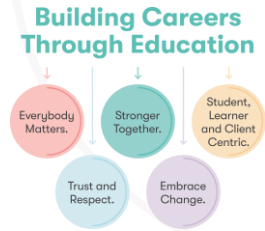
Optimisation: Rewriting

- Simplify complex queries to reduce execution time.
- Use subqueries, joins, and unions efficiently.
- Example: Replace correlated subqueries with joins.



BPP

More optimisation ideas



Denormalisation

- In some cases, denormalise data to reduce the number of joins.
- Trade-offs between storage space and query speed.

Partitioning

- Split large tables into smaller, more manageable pieces.
- Improve query performance by limiting the amount of data scanned.

Caching

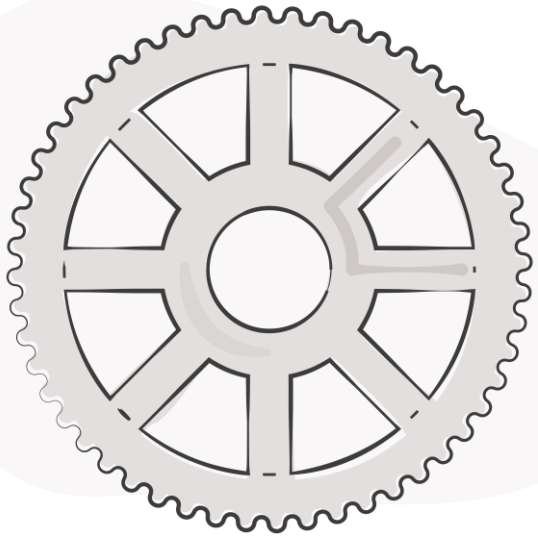
- Use caching mechanisms to store frequently accessed data in memory.
- Reduce the need for repeated database access.

BPP

Configuration tuning

- Adjust database configuration settings (e.g., memory allocation, buffer sizes).

- Tailor settings to match workload characteristics and hardware resources.



Database recovery models

- A recovery model is a database property that controls how transactions are logged, whether the transaction log requires (and allows) backing up, and what kinds of restore operations are available.

- Typically, a database uses the full recovery model or simple recovery model. A database can be switched to another recovery model at any time.

- Simple = only the latest backup can be restored.

- Full = can restore to different points before failure.

BPP

PITR

- PITR (point in time recovery) combines elements of full recovery with automation to ensure data integrity and availability

- Using automated snapshots, allows restoring to any specific time.

- In AWS, PITR offloads storage and maintenance to AWS, not requiring local management.

BPP

Amazon RDS Snapshots

Amazon RDS snapshots are critical for efficient database management and disaster recovery. These snapshots capture the state of your database at a specific point in time, allowing you to restore your database to that exact state when needed.

Snapshot types:

- **Automated Snapshots**: Triggered daily within a set backup window. Ideal for regular, scheduled backups.
- **Manual Snapshots**: User-initiated anytime for on-demand backups.

You pay only for the incremental storage used.

BPP

Indexing

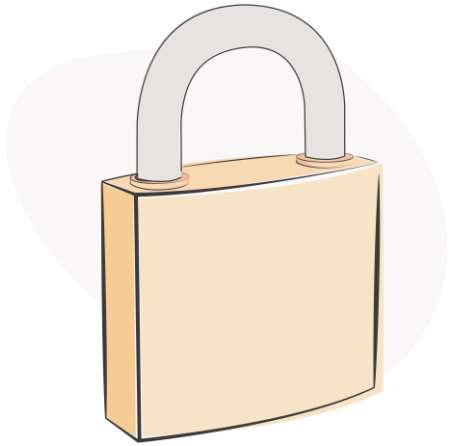
EXTREMELY IMPORTANT

- If you're not indexing, you're not *close* to full performance
- "Covered" queries
- Clustered versus not…
- Examine your `SELECT` statements
- Order in the `SELECT`

BPP

Security

- Don’t have all developers login with the same UID
- Don’t use the DBA UIDs for developers
- Do keep UIDs confidential
- Don’t let an application use a developer/user/dba login



SQL Injection

Can provide alarming access

Select * from myTable where lname=‘” request.form(“lastname”) & “’”

Form input ’ or ‘a’=‘a

Result: `select * from mytable where lname='' or 'a'='a`

How to protect against it

- Stored procedures
- Watching quotes
- Control statement generation
- Parameter queries

BPP

Outages

Take a guess

Which of the following do you think the leading cause of datacenter outages?

A) Power outage

B) Over-heating

C) Human error

D) Fire

E) DOS attacks

Feedback: C – Human error (70%)

Submit your responses to the chat!

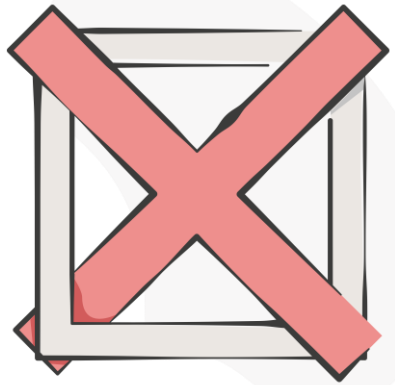
Outages

- A system operator mistakenly deleted the $38 billion Alaska Permanent Fund database and then deleted its backup.
- A maintenance contractor’s mistake shut down the Oakland Air Traffic Control Center.
- A State of Virginia technician pulled the wrong controller and crashed a redundant SAN that already had suffered a controller failure.
- A technician with DBS Bank made an unauthorized repair on a redundant SAN and took down both sides.

Source:
http://www.availabilitydigest.com/public_articles/0704/data_center_outages-lessons.pdf

Outages

- Outages are inevitable
- AWS, Facebook, other big companies have suffered massive outages but they always kept affected users updated throughout
  - Frequent updates
  - Coupons/discounts
  - Published post-mortems afterwards
  - All these bolster customer confidence
- Many companies run dashboards with real-time information
  - Google Apps status dashboard
  - AWS dashboard



Exercise

Learning Summary

- Data profiling
- Database archiving
- Query profiling
- Query optimisation
- Recovery
- Indexing
- Security
- Outages

Thank you

Do you have any questions, comments, or feedback?