Data Archiving and Purging:
Concepts, Approaches, and Technologies

NWA Technology Conference September 18, 2019

DXC.technology

Why are we even talking about this?

A long time ago in a galaxy far, far, away…

Do we need a data archiving and purging strategy?
“Hey, disk space is cheap we’ll worry about that later!”

Present day…

- Systems are bogged down with historical data.
- We need faster servers and more storage capacity to handle the ever-growing data requirements of the system.
- Database housekeeping and backups are colliding with nightly processing.

Why is Archiving and Purging Important?

- **Quality**
  - Keeps relevant information at the fingertips of the business
  - Allows for access to historical records that must be kept
  - Destroys data that is no longer needed
- **Time**
  - Optimizes access to relevant operational data
  - Reduces duration of house keeping efforts. (E.g. backups, re-indexing, resizing, etc.)
- **Costs**
  - Reduces the need to invest in larger, faster resources to handle ever growing data stores.
  - Allows for lower-cost long-term storage options

Do I really need to be worried about this?

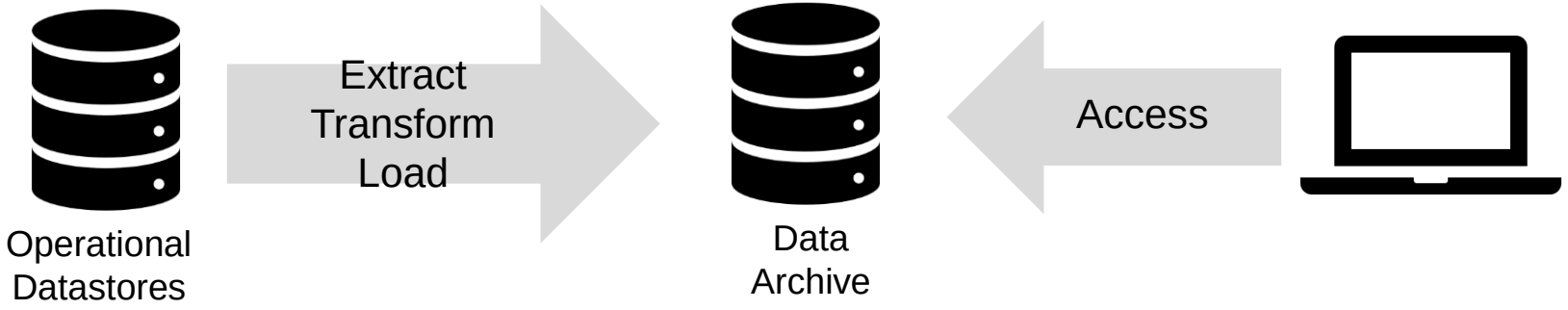
- Is your system “bogged down” with information that is no longer relevant?
- Are you seeing a degradation in the performance of your system?
- Are you being told that you need a bigger/faster database server and you don’t want to buy a new one?
- Are your data storage costs becoming to high or are you running out of storage space and you don’t want to buy more?
- Is your IT group raising concerns over long-running maintenance tasks and/or database backups that are encroaching on your nightly processing?

If you answered yes to any of these questions, then you might want to consider implementing a data archiving and purging strategy.

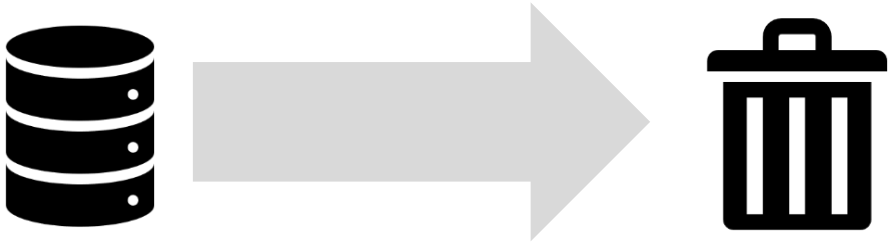
Concepts

Archiving vs. Purging

Archiving - Moving data from one datastore source to another.



Purging is the process of deleting data from a datastore.



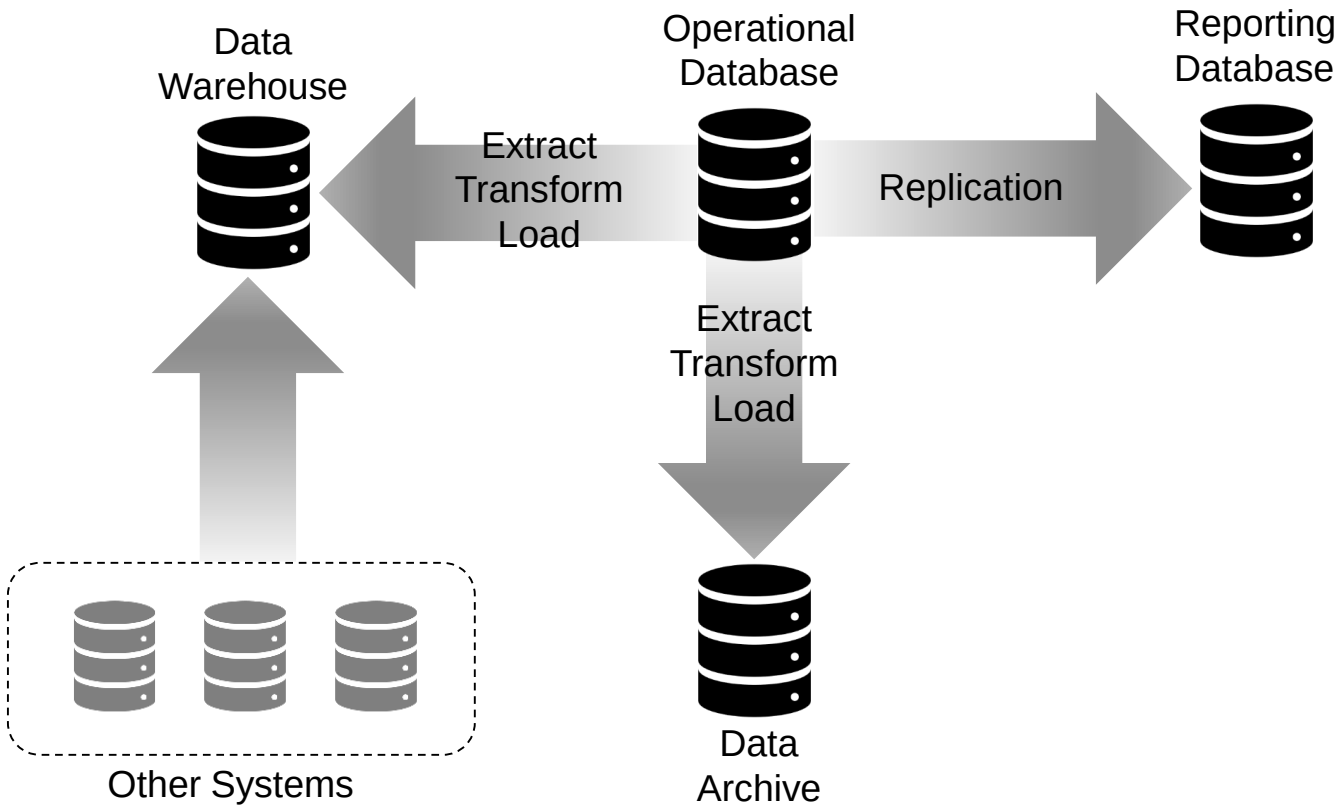
Data Retention Policies

- Defines how long you want to keep information.
- An effective data retention policy meets legal, regulatory, and business requirements.
- Data retention policies can be driven by business needs as well as technical constraints.
- Data retention policies can be applied to both archiving and purging.

Types of datastores

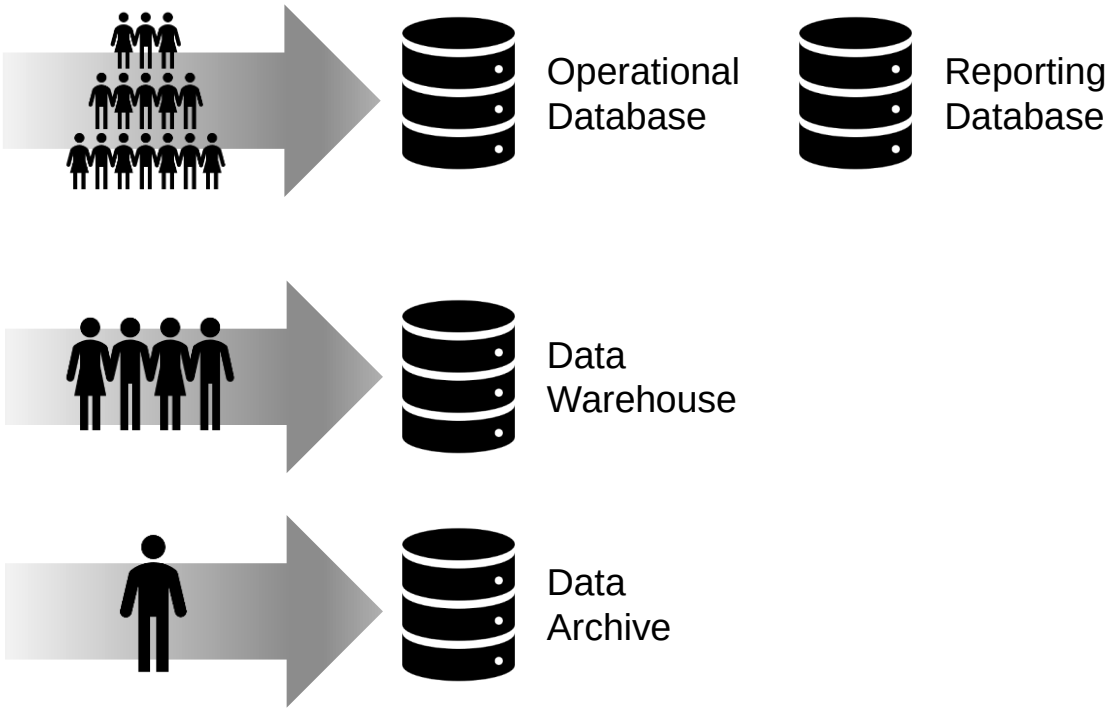
- **Operational Datastores**
  - These are the “main” datastores for the system. They can include **databases** and **file systems**.
- **Reporting Database**
  - This is typically a “copy” of the operational database used as a data source for generating operational reports.
  - This database is typically populated using a replication strategy; transactional replication, database mirroring, log shipping, nighty backup and restore, etc.
- **Data Warehouse**
  - This is typically a datastore that has been optimized to provide reporting capabilities and data analysis beyond the operational reporting database.
  - It is **not** a replicated copy of the operational database.
  - Typically includes data from other data sources.
  - Data is periodically Extracted (**copied**) from the operational database, Transformed, and Loaded into the Data Warehouse.
- **Data Archive**
  - This is typically a long-term records management solution for historical data.
  - Access is typically infrequent and limited to small group of individuals.
  - Data is periodically Extracted (**moved**) from the operational database, Transformed, and Loaded into the Data Archive.
- **Data Backup**
  - Long term off-line storage used for data recovery purposes.

Datastore Relationships

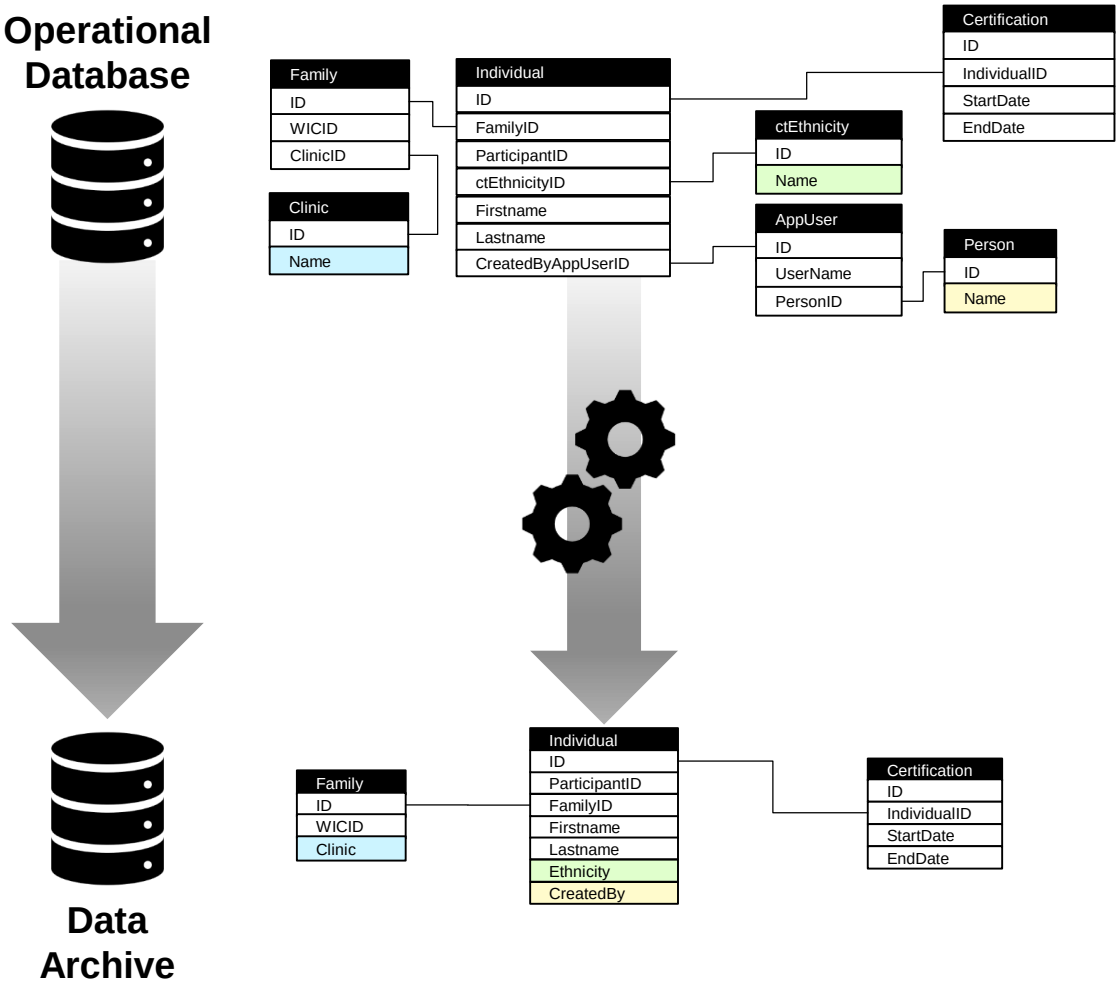


Who needs to get to what data?

- **Day-to-day operations**
  - Participant Services
  - Vendor Management
  - Finance
  - Operations and Administration
  - Real-time reporting
  - Operational reporting
- **Data Analytics**
  - Trending and forecasting
  - Historical reporting
  - Spatial analysis
  - Enterprise reporting
- **Historical document retrieval**
  - Litigation
  - Inquiries
  - Research



Extract, Transform, and Load



Extract

Identify and remove the data to be archived.

Transform

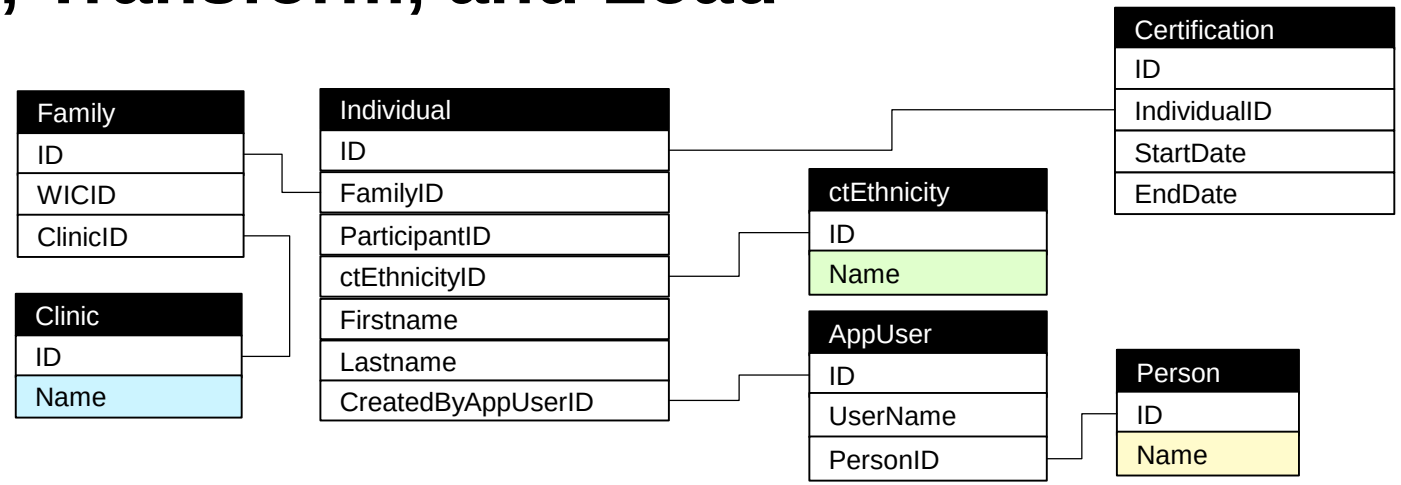
Re-shape the data into a form more suitable for the archive.

Load

Insert the transformed data into the archive.

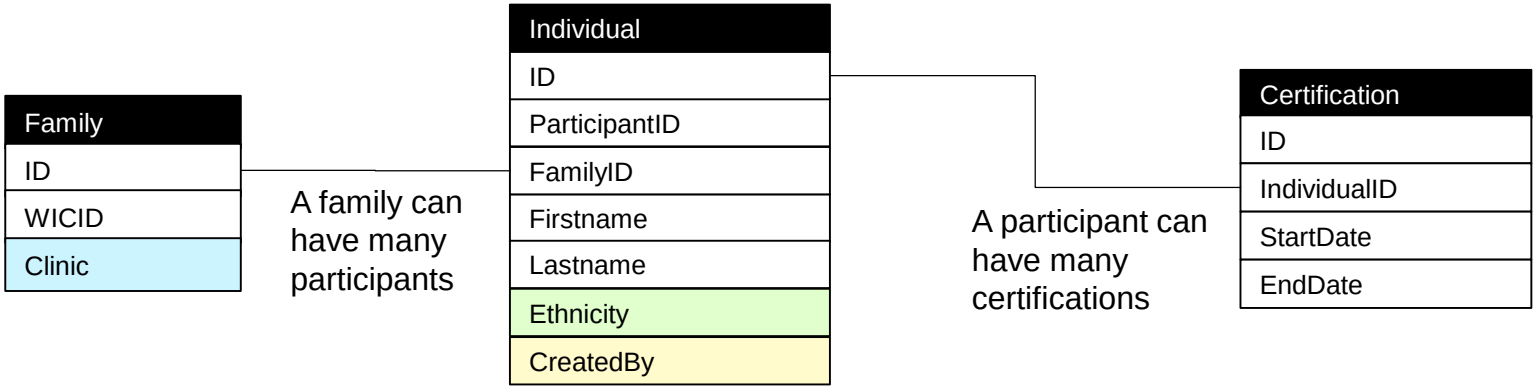
Extract, Transform, and Load

Operational Database



- Simplified (de-normalized) the structure
- Translated and captured the actual code values at time the data was archived
- Preserved one-to-many parent-child relationships

Data Archive



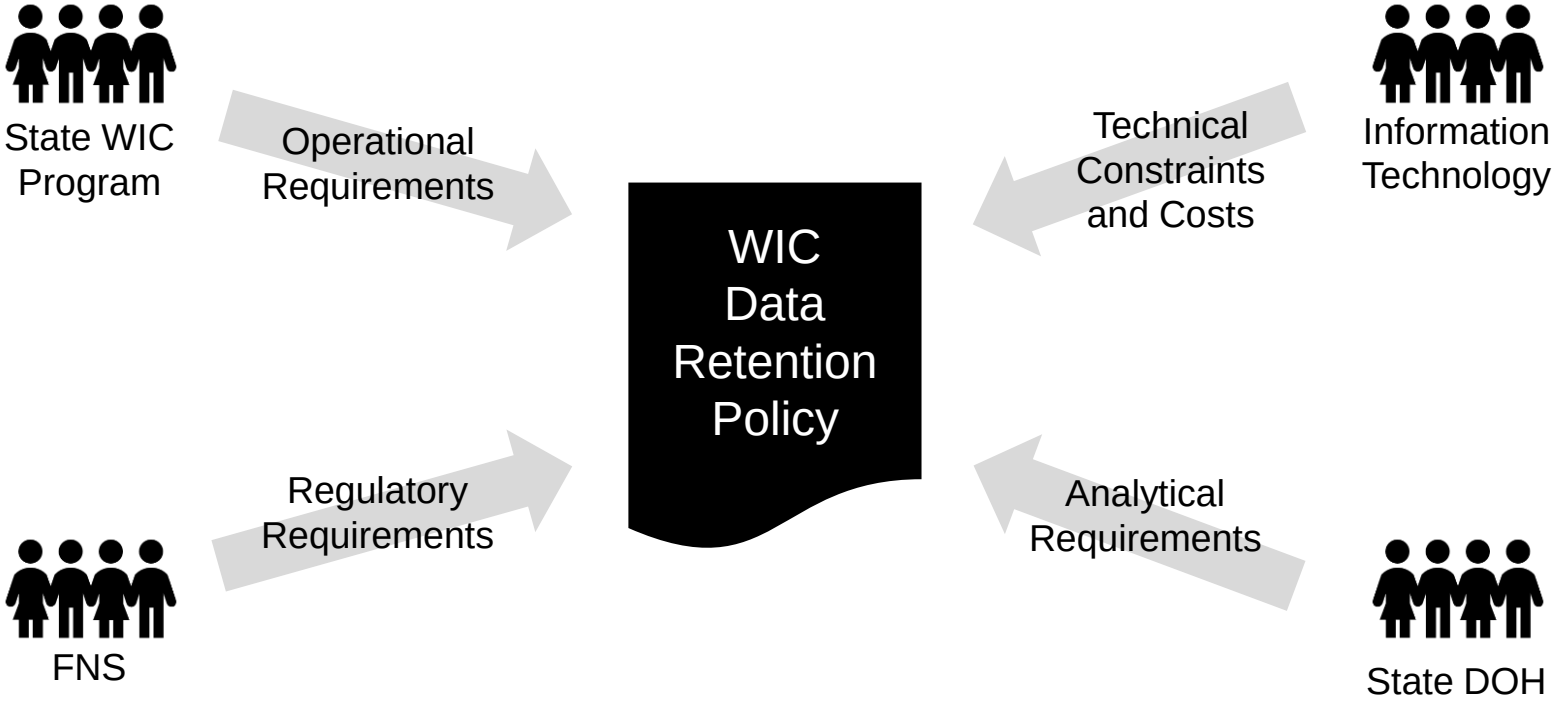
Approaches

Approaches

- **Keep all the data forever. (Do not implement an archiving and purging strategy)**
  - Scale up (add memory and/or CPU to your servers), scale out (add more servers), or replace your servers as needed.
  - Use database features to optimize performance. E.g. Database Partitioning
  - Add more disk space as needed.
  - Adjust (as best you can) backup plans to avoid overlapping with nightly processes.
- **Implement a Data Archiving and Purging strategy.**
  - Archiving moves the old data, that must be retained, out of the operational datastores to the data archive.
  - Purging deletes data from the operational datastores and data archive that does not need to be retained.
  - Implementing both archiving and purging provides the best chance to improve the performance of your operational database.
- **Only implement a Purging strategy.**
  - Purging deletes data from the operational datastores and data archive that does not need to be retained.
  - Avoids the complexities and cost of building and maintaining a Data Archive.
  - Leaves old data, that is not used, but must be retained, in the operational datastores eliminating any performance improvements that might have been gained by archiving.

Identify your Data Retention Policies

Get input from all the stakeholders



Data Retention Policies

- **Federal Regulations, 7 C.F.R. Part 246.25 Records and reports.**
  - (1) Records shall include, but not be limited to, information pertaining to financial operations, food delivery systems, food instrument issuance and redemption, equipment purchases and inventory, certification, nutrition education, including breastfeeding promotion and support, civil rights and fair hearing procedures.
  - (2) **All records shall be retained for a minimum of three years** following the date of submission of the final expenditure report for the period to which the report pertains. **If any litigation, claim, negotiation, audit or other action involving the records has been started before the end of the three-year period, the records shall be kept until all issues are resolved, or until the end of the regular three-year period, whichever is later.** If FNS deems any of the Program records to be of historical interest, it may require the State or local agency to forward such records to FNS whenever either agency is disposing of them.
- **State WIC Program requirements may go beyond the Federal Regulations.**

Data Retention Policies: Sample

Based on Federal Regulations, 7 C.F.R. Part 246.25

- **Participant Services**
  - Archive families that have been inactive* for over 24 months that are not flagged to be retained in the operational database. When archiving redemption data associated with a family, if the redemptions are associated with a vendor **copy** the family's redemption data to the archive otherwise **move** the redemption data.
  - Purge families that have been archived for over 12 months.
- **Vendor Management**
  - Archive vendors that have been terminated or disqualified for over 24 months that are not flagged to be retained in the operational database. When archiving redemption data associated with a vendor, if the redemptions are associated with a family **copy** the vendor's redemption data to the archive otherwise **move** the data.
  - Purge vendors that have been archived for over 12 months.
  - Purge all file logs older than 6 months.

* For purposes of this presentation, a family is considered inactive when there are no active certifications for any participants in that family.

Data Retention Policies: Sample

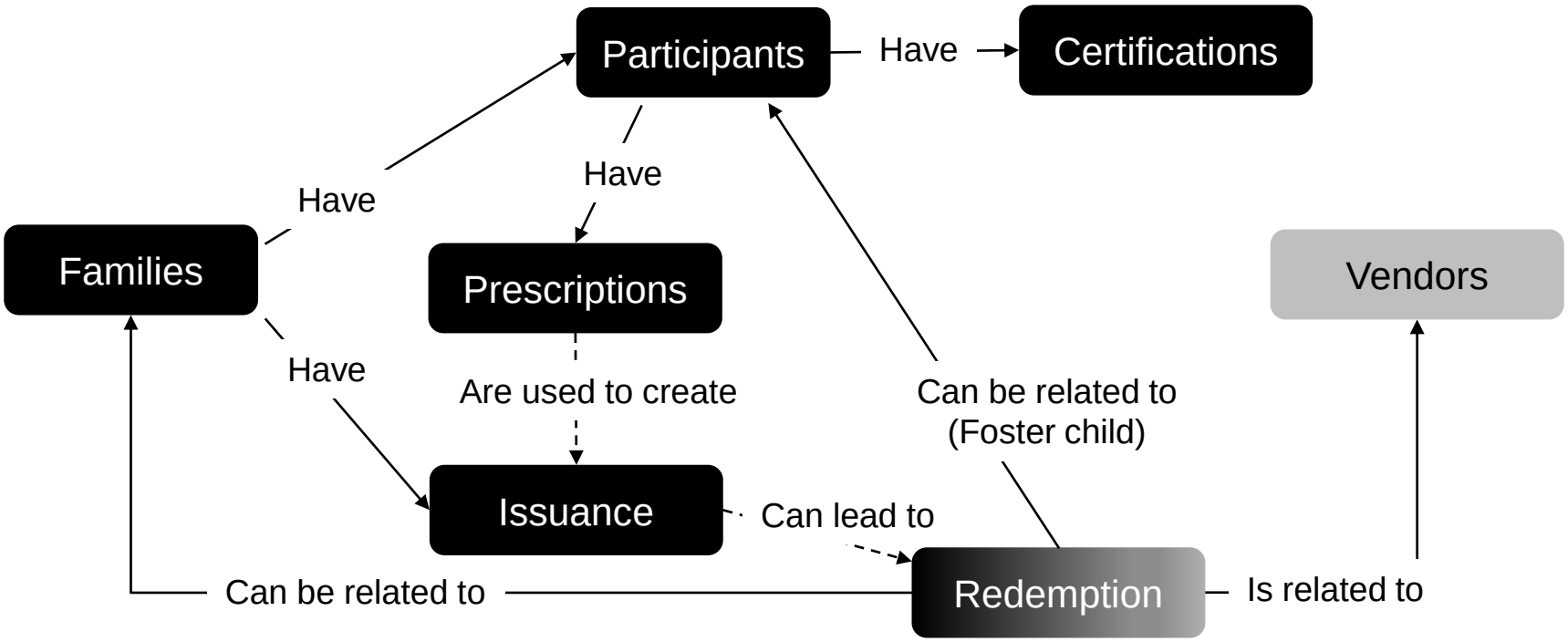
Based on Federal Regulations, 7 C.F.R. Part 246.25

- **Finance**
  - Archive months that have been closed for over 24 months.
  - Purge months that have been archived of over 12 months.
- **Audit History**
  - Purge records older than 6 months*.
- **System Logs**
  - Purge message logs older than 3 months*.

* For purposes of this presentation, assumes that 7 C.F.R. Part 246.25 does not apply to this data.

Data Relationships: Example

Archiving and Purging at a Family level keeps all the related data together.



Redemption data related to a family may need to be copied to the archive if it is related to other entities in the operational database.

Archiving: Example

- Current Database size is 135 GB.
- Number of active certifications is 133,000.
- Database has approximately 1 million participants grouped into 430,000 families.
- **79% of those families have been inactive* for over 24 months**. Archiving these families would result in an estimated 61 GB reduction in the overall size of the database.
- **72% of those families have been inactive* for over 36 months**. Archiving these families would result in an estimated 52 GB reduction in the overall size of the database.
- **54% of those families have been inactive* for over 48 months**. Archiving these families would result in an estimated 43 GB reduction in the overall size of the database.

* For purposes of this presentation, a family is considered inactive when there are no active certifications for any participants in that family.

Archiving: Example

			24 Months			36 months			48 months		
Table	Row Count	Table Size (GB)	Reduction	Row Count	Table Size (GB)	Reduction	Row Count	Table Size (GB)	Reduction	Row Count	Table Size (GB)
Individual	1,066,331	0.83	71%	309,293	0.24	62%	400,358	0.31	54%	493,150	0.38
Family	429,690	0.20	79%	88,274	0.04	72%	119,907	0.06	64%	153,847	0.07
FoodInstrument	20,257,889	25.16	89%	2,156,969	2.77	83%	3,421,093	4.28	75%	5,015,176	6.29
FoodInstrumentDetail	78,626,075	10.81	81%	14,567,162	2.05	72%	21,738,805	3.03	62%	29,866,067	4.11
FoodPrescription	3,703,615	0.42	49%	1,880,537	0.21	37%	2,350,455	0.26	24%	2,803,245	0.32
FoodPrescriptionDetail	21,260,259	2.15	54%	9,756,955	0.99	41%	12,440,491	1.27	29%	15,189,385	1.53
Certification	2,215,525	1.33	65%	774,664	0.46	54%	1,022,341	0.61	42%	1,289,509	0.77
Appointment	3,189,065	2.14	38%	1,983,781	1.32	22%	2,477,276	1.67	10%	2,874,519	1.92
EBTRedemptionTransaction	16,199,964	6.98	72%	4,597,011	1.96	58%	6,766,365	2.93	44%	9,057,139	3.91
EBTRedemptionTransactionDetail	65,229,847	25.94	71%	19,126,665	7.52	57%	27,874,526	11.16	43%	37,002,847	14.79
EBTBenefitActivity	3,770,105	0.68	75%	946,878	0.17	61%	1,479,999	0.27	47%	1,981,850	0.36
EBTBenefitActivityDetail	28,983,746	2.20	75%	7,172,121	0.55	61%	11,221,224	0.86	48%	15,143,198	1.14

Purging: Example

- Current Database size is 135 GB.
- 9 GB is audit history.
- Audit history is a “before image snapshot of a record that has been changed”.
- On average, 75% of this audit history data is more than 6 months old. Purging this data would reduce to overall size of the database by 6.8 GB.
- If you don’t want to “lose” this data than archive it and purge it from the archive later!

Considerations

- Is your Data Analytics team using the Reporting Database as the data source for analytics? If so, archiving and purging your operational database may impact the Data Analytics team.
- Who should have access to the archive? What training will they need.
- What should the outputs from the archive look like?
  - Are they raw (adhoc) query results?
  - Are they formatted reports or pdf documents?
- Work with your M&E Vendor to build your archiving and purging processes.
  - Each archiving and purging implementation will vary based on the system.
  - Considerations must be given to explicit data relationships **as well as** implicit data relationships.

Technologies

Technologies

Here a few technologies to consider. This is not a complete listing.

- **ETL tools**
  - Microsoft SQL Server Integration Services
  - IBM Infosphere DataStage
  - Oracle Data Integrator
  - Talend
  - Informatica
- **Archive repositories**
  - SQL Server
  - Oracle
  - MySQL
  - MongoDB
  - File System
- **Archive location**
  - On Premise
  - Amazon S3 Glacier
  - Azure Cool Blob Storage

About DXC Technology

DXC Technology has been providing WIC solutions for over 40 years.

As the world's leading independent, end-to-end IT services company, DXC Technology (NYSE: DXC) leads digital transformations for clients by modernizing and integrating their mainstream IT, and by deploying digital solutions at scale to produce better business outcomes. The company's technology independence, global talent, and extensive partner network enable 6,000 private and public-sector clients in 70 countries to thrive on change. DXC is a recognized leader in corporate responsibility. For more information, visit dxc.technology and explore thrive.dxc.technology, DXC's digital destination for changemakers and innovators.