An Overview of Architectures and Techniques for IDS Implementation

Prashant Kumar

Managing Partner, Integrating Factors Inc. (IFI)

www.ifipublicsector.com

Discussion Points

- Why a integrated data system/platform is needed
- Bringing the data together (data architecture)
  - Federated data, data warehouses, data marts, operational data stores
- Matching / linking of identities across systems (clients and providers)
- Integrating data (Integration approaches/architecture)
  - Need based
  - Periodic
  - Continuous
- Retrieving data
  - Search
  - Data Delivery (BI, stat analysis, data mining)
- Data governance
- Enabling technologies / tools

Why integrated data system/platform is needed

- Program orientation of HHS systems
- Programmatic funding sources

Program-orientation of data systems = Lack of an end-to-end client-centric view

Practice

- Risks, care gaps
- Service coordination

Policy

- Important patterns are obscured;
- Chain of consequences/factors harder to understand

Why integrated data system/platform is needed …

- Program-oriented HHS systems address the operational/administrative needs
- But, often fall short in meeting informational needs for problem solving
- IDS can bridge the gap

Applications of IDS

Case Data

- Assessment
- Service Coordination
- Cross-Agency Alerts
- …

Population Data

- Policy Analysis
- Target Population Identification and Stratification
- Utilization Analysis
- …

Bringing the data together (Data Architecture)

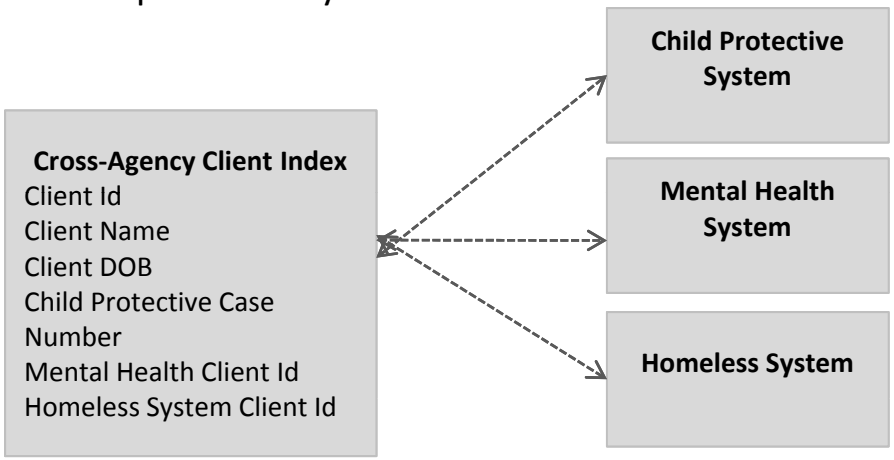
Federated Data
Data Warehouse
Operational Data Store

Bringing the data together (Data Architecture)

- Federated Data
- Data Warehouse
- Operational Data Store
- Hybrid

Federated Data

- Data stays in the transactional data systems
- A cross-reference index maps data to systems



- Query decomposition through shared cross-reference data
- Dynamically extracted and presented to the user or application

Advantages

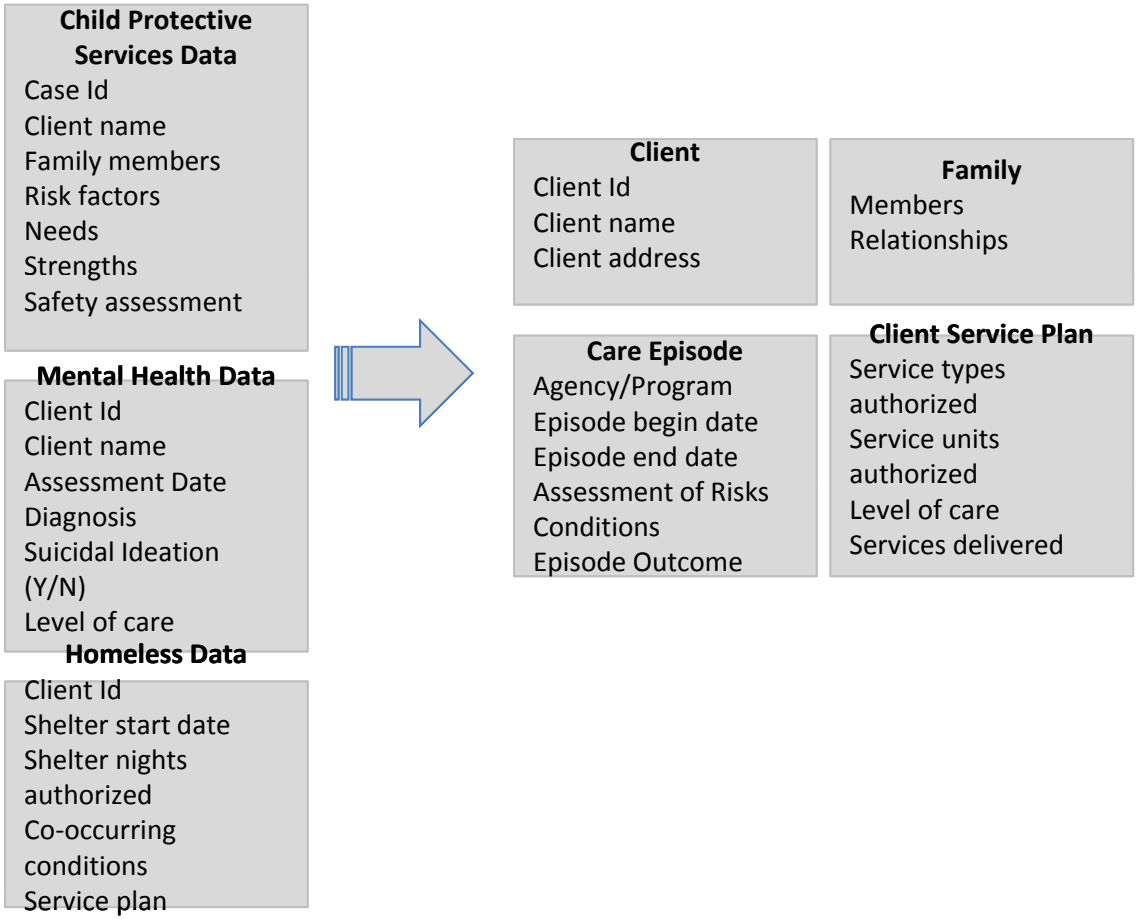
- Legal compliance –> data visibility
- Usability –> programmatic context

Disadvantages

- Full availability = that of least available source

Data Warehouse

- Shared, integrated database
- Subject oriented, integrated, non-volatile, historical
- Structure – Inmon Vs. Kimball approaches



- Ideal for decision support functions, analytics, Data quality improvements, models
- Latency, Scope/complexity/cost

Operational Data Store

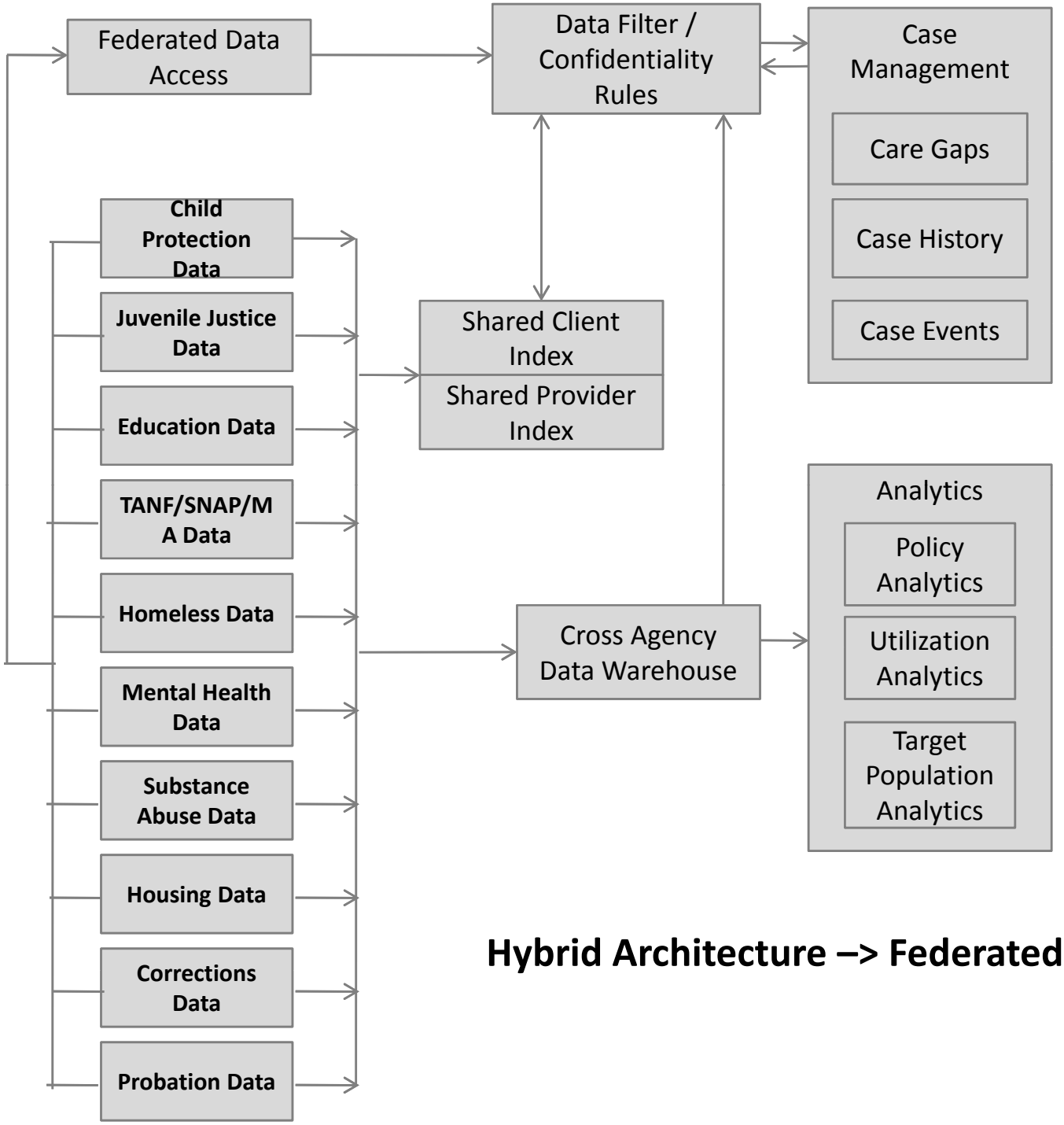
- Shared, integrated database
- Subject oriented, integrated, **volatile and current**

Advantages:

- Ideal for case level analytics
- Data quality improvements, models

Disadvantages:

- Scope/complexity/cost



Hybrid Architecture –> Federated & DW

Matching / Linking Identities – clients and providers

- Model type - deterministic or probabilistic
- Weight Assignment
- Match thresholds
- Blocking approaches
- Link cascades
- Adhesion factor of linked records

Model Type

- Deterministic model parameters estimated a priori
- Probabilistic model parameters estimated based on data
- History
  - Newcombe
  - Fellegi & Sunter
  - Many practical applications by Winkler and others

Fellegi Sunter – Key Idea

$$Likelihood\_Ratio = \frac{P(comparison\_pattern \mid Match)}{P(comparison\_pattern \mid NonMatch)}$$

R >= some upper threshold = link
R <= some lower threshold = non-link
Lower threshold < R < upper threshold = possible link, manual review needed

Type I and Type II errors depend on thresholds

Example:

$$Likelihood\_Ratio = \frac{P(agree\_SSN \ \& \ LastName \mid Match)}{P(agree\_SSN \ \& \ LastName \mid NonMatch)}$$

Fellegi Sunter – assumptions

With conditional independence assumption:

$$R = \frac{P(agree\_SSN \mid Match)}{P(agree\_SSN \mid NonMatch)} + \frac{P(agree\_LASTNAME \mid Match)}{P(agree\_LASTNAME \mid NonMatch)}$$

Using conventional notation:

$$R = \frac{m(SSN)}{u(SSN)} + \frac{m(LASTNAME)}{u(LASTNAME)}$$

Fellegi Sunter …

$$\log R = \log \frac{m(SSN)}{u(SSN)} + \log \frac{m(LASTNAME)}{u(LASTNAME)}$$

In general:

$$\log R = \sum \{\log \frac{m}{u} \; forAgreements, \log \frac{1-m}{1-u} \; forDisagreements\}$$

Fellegi Sunter … example

- Training dataset has following m and u probabilities:
  - SSN m=0.95, u=.08
  - Last Name m= 0.80, u=0.15
- A pair of records agree on both SSN and Last Name

$$\log R = \log \frac{0.95}{0.08} + \log \frac{0.8}{0.15}$$

$$\log R = 3.57 + 2.42 = 5.99$$

Fellegi Sunter … another example

- A pair of records agree on SSN but disagree on Last Name

$$\log R = \log \frac{0.95}{0.08} + \log \frac{1-0.8}{1-0.15}$$

$$\log R = 3.57 - 2.09 = 1.48$$

Comparisons

- Agreements
- Disagreements
- Partial agreements
  - Edit distance measures

Matching Steps

- Estimate m and u probabilities for each comparison
  - Training dataset / prior experience / iterate
  - EM and ECM
- Calculate weight of agreement/disagreement for each comparison

$$agreement\_weight = \log\frac{m}{u}$$

$$disagreement\_weight = \log\frac{1-m}{1-u}$$

- Ascertain Type I and II error tolerance
  - Legal / compliance implications
  - Conditional independence assumption = understated error rate projections
- Set thresholds
- Match by calculating R = Sum of weights for agreements and disagreements
- Empirical approach – capture-recapture / eyeball
- Iterate

Other Matching / Linking Considerations

- Blocking
- Link cascades
- Adhesion factor
- Persistent manual overrides

Blocking

- Computational cost of all possible matches
- Typical blocking variables:
  - Phonetic (Soundex, NYSIIS)
  - First n characters of last name
  - …
- Type II misclassification with blocking
- Multiple passes to minimize Type II

Link Cascade

Record #	Source	Last Name	First Name	Middle Name	SSN	Birth Date	Matching Record	Linked Record
1	Corrections	Webb	Mary	J	515433219	1/02/72	2	2, 3
2	Child Welfare	Jones	Mary		515433219	1/22/72	1, 3	1, 3
3	Homeless	Jones	Mary		515433291	1/22/72	2	1, 2

Adhesion Factor Over Time

	Last Name	First Name	Middle Name	Address	SSN	Birth Date	Matching Record	Linked Record
	Webb	Mary	J	123 Main St., Any City, USA	515433219		2	2, 3
	Webb	Mary		123 Main St., Any City, USA		1/22/72	1	1, 3
	Jones	Mary			515433291	1/22/72	None	1, 2

Integration Approach

- Need Based
- Periodic
- Continuous

Needs Based Integration

- Addresses specific analytic need or business issue
- One-off project

Advantages

- Get started with cross-agency data sharing and coordination
  - governance structure, work with peer agencies
- Quick return

Disadvantages:

- One-off effort;
  - Legal reviews
  - Data preparation
  - Lack of systematic process controls for confidentiality management
- Not for case level work

Periodic Data Integration

- Data integration process
- Pre-determined time intervals – quarterly/annually etc.

Advantages

- Ready to use data
- Programmatic and legal reviews obtained once;
- Emphasis on data integration as a process – process maturity over time.
- Window for data reconciliation and reasonability testing

Disadvantages:

- Not suitable for case level work due to the inherent latency of the periodic data integration process.

Continuous Integration

- Real-time/near real-time
- Process-driven – process maturity on day 1

Advantages

- Ready to use data
- Suitable for case level work

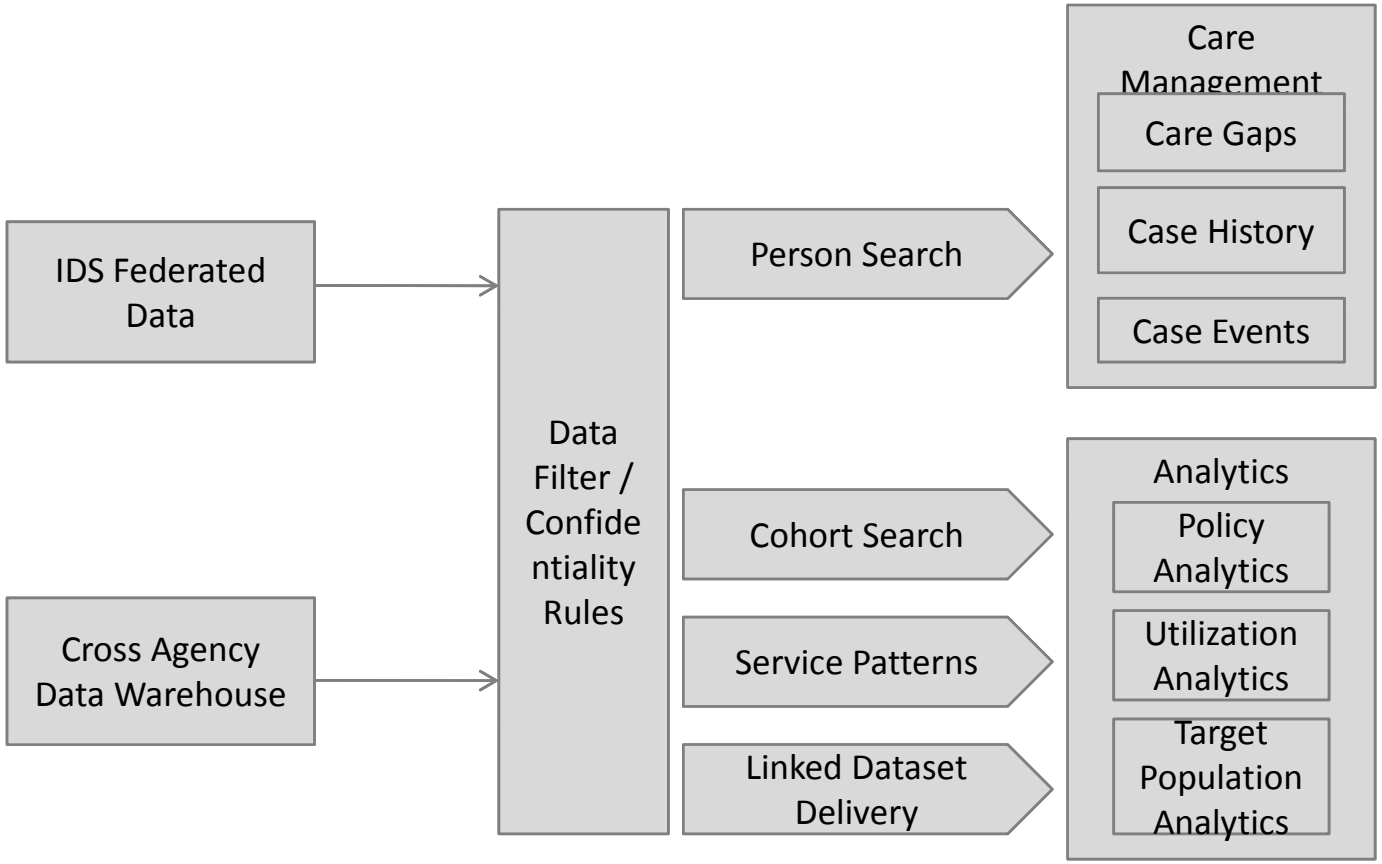
Disadvantages:

- No window for manual reviews, reasonableness testing etc.
- If deployed for case level work, complexities of client consent, waiver, court orders

Information Retrieval

- Search
- Data Delivery

Information Retrieval - Search



- Find a client with natural identifiers or agency identifiers;Address “First Visit Issues”
- Find common clients by matching client files/lists
- Find clients with a given service pattern – “Show me all clients with ….”

Information Retrieval - Delivery

- Confidentiality rules
  - Purpose
  - Ontology – what information (category), for whom (agency, provider, policy analyst, researcher, etc.)
  - Consent (opt-in, opt-out, category, participant)
  - Waiver
  - Court order
  - Logging/audit trail
- Linked data delivery
  - Portal-based
  - BPM-based
  - Longitudinal / flat records for data mining/stats packages

Enabling Technologies

- Data exchange
- Data integration
- Information delivery

Enabling Technologies

Data Exchange

- File transfer
- Message oriented middleware
- Connection oriented middleware
- EAI suites
- Web services

Enabling Technologies

Data Integration

- Data profiling
- Data cleansing
- Identity matching
- ETL
- Data replication
- Change data capture

Enabling Technologies

Information Delivery

- Portals
- BPM (Business Process Management)
- BI (Business Intelligence)
- Stats analysis / data mining

Data Governance

- **Data Use**
  - Data use policies
- **Legal Risk**
  - Data sharing agreements
  - Monitoring/control
- **Data Quality** –
  - monitoring, measuring and controlling the accuracy, timeliness, consistency and reliability of data.
- **Technology Architecture and Data Exchange Standards**
- **Service Levels** – Clarify roles and responsibilities of data provider and data consumer agencies.

Cost Factors

- Number of data sources
- Complexity of data sources
- Data quality levels
- The number and complexity of use cases
- Legal requirements for data categories to be shared
- Technology platform of programmatic data systems
- IDS architectural approach
- ….

Q&A