P2-FED: Methodology & Quality Assurance

1. Data Source & API Endpoints

Primary Source: USAspending.gov - Official source of federal spending data mandated by the Digital Accountability and Transparency Act (DATA Act).

API Endpoints Used

```
POST /api/v2/search/spending_by_award/
```

Purpose: Retrieve award-level spending data with pagination support

```
POST /api/v2/search/spending_by_award_count/
```

Purpose: Get total record counts for scope estimation

Request Parameters

Parameter	Value	Description
time_period	Multi-year (FY2004-2027)	Includes multi-year contract obligations
agencies	DoD, HHS, DHS	Top-tier awarding agencies
award_type_codes	A, B, C, D	Contract types: BPA, PO, DO, Definitive
limit	100	Records per API call

Note: Federal contracts often span multiple fiscal years. Total spend reflects cumulative obligations across all contract periods.

2. Data Model (Star Schema)

Fact Table: award_fact

Column	Type	Description
award_id	VARCHAR	Unique award identifier
award_amount	DECIMAL	Total obligation amount (USD)
agency_id	INT	FK → agency_dim
recipient_id	INT	FK → recipient_dim
start_date	DATE	Award period start
fiscal_year	INT	Derived fiscal year
naics_code	VARCHAR	Industry classification
psc_code	VARCHAR	Product/Service code

Dimension Tables

Table	Key Columns	Row Count
agency_dim	agency_id, toptier_name, subtier_name	6
recipient_dim	recipient_id, recipient_name, uei	7452
time_dim	date_id, date, fiscal_year, quarter	2160
geo_dim	geo_id, state_code, city_name	58

3. Quality Gates

Gate	Threshold	Logic
Schema Drift	≥95%	Verify expected columns present; null rates < 1% for key fields
Freshness	≥80%	Latest record date within 365 days; penalize stale data
Completeness	≥90%	Null rates for agency, recipient, amount < 5%
Duplicates	≥95%	Hash-based deduplication on award_id + recipient + amount
Value Sanity	≥85%	Negative amounts < 5%; z-score outliers < 1%
Referential Integrity	≥95%	FK references exist in dimension tables

Quality Score Formula

```
Overall = Σ(gate_score × weight) / Σ(weights)

Weights: schema=0.15, fresh=0.15, complete=0.20, dupe=0.15, sanity=0.20, RI=0.15
```

4. KPI Definitions

Spend Metrics

Metric	Formula
Total Spend	SUM(award_amount) by dimension
Average Award	AVG(award_amount)
Median Award	PERCENTILE_CONT(0.5) of award_amount

Concentration Metrics

```
Herfindahl-Hirschman Index (HHI):
HHI = Σ(market_share²) × 10,000

where market_share = vendor_spend / total_spend

Interpretation:
• < 1,500: Unconcentrated (healthy competition)
• 1,500-2,500: Moderately Concentrated
• > 2,500: Highly Concentrated (potential risk)
```

Change Detection

Metric	Formula
QoQ Change	(Q_current - Q_prior) / Q_prior × 100%
Rank Change	Vendor rank position delta between periods

5. Pipeline Architecture

Stage	Description	Output
1. Ingest	Paginated API calls with cursor-based pagination	raw_awards_fy2024.csv
2. Clean	Type conversion, null handling, deduplication	cleaned_awards_fy2024.csv
3. Model	Star schema transformation, FK generation	award_fact.csv, *_dim.csv
4. Validate	Quality gate execution	pipeline_metrics.json
5. Analyze	KPI calculation	kpis.json
6. Report	PDF generation	Executive + Methodology PDFs

6. Run Metrics

Metric	Value
Pipeline Start	2026-02-01T06:47:28.006184+00:00
Pipeline End	2026-02-01T06:50:03.992831+00:00
Duration	156.0 seconds
API Calls	248
API Errors	11
Raw Records	24,800
Cleaned Records	24,800
Modeled Records	24,800
Overall Quality Score	96.7%

Report prepared by Mboya Jeffers | MboyaJeffers9@gmail.com

Data Source: USAspending.gov API | Documentation: api.usaspending.gov/docs/endpoints

Generated: 2026-02-01 07:19 UTC