TECHNICAL AUDIT & STRATEGIC ANALYSIS

TMDb Data Engineering Master Documentation

A comprehensive audit of the pipeline architecture, data integrity controls, and capital intelligence derived from the TMDb 5000 dataset.

---

01 METADATA ARCHITECTURE

Audit of data ingestion, “Mojibake” text correction, and financial verification.

02 JSON PARSING & NORMALIZATION

Documentation of Python migration, Crew Segmentation, and Relational Modeling.

03 STRATEGIC INTELLIGENCE

Capital allocation heuristics: Budget Strategy, Risk Profiling, and Consistency Analysis.

ANALYST

Mark Anthony O. Nene

DATE

August 2025

TECH STACK

Excel • Power Query • Python

STATUS

✓ **Master Audit Locked**

How to Use This Document

> **Document Purpose:** To audit the technical methodology, cleaning logic, and validation steps used to transform the raw TMDb dataset into a relational asset.

Intended Audience & Navigation

This document is designed for **non-linear reading**. Please navigate to the section that aligns with your specific audit criteria:

1. For Analytics Leads & Hiring Managers (Strategic Validity)

- **Focus:** How were business conclusions derived?
- **Jump to:**
  - **Director Risk Profiling:** Review Capital Efficiency modeling.
  - **Actor Consistency Analysis:** Review Risk-Adjusted Return metrics.
  - **Financial Verification:** See “Preservation over Deletion” logic.

2. For Data Engineers & Technical Reviewers (Pipeline Integrity)

- **Focus:** Is the code robust and scalable?
- **Jump to:**
  - **JSON Architecture:** Review Python migration for nested arrays.
  - **Crew segmentation:** See Boolean filtering strategy.
  - **Validation Logic:** Review “Financial Density” checks.

Technical Stack

- **Extraction & Parsing:** Python (Pandas, JSON)
- **Transformation & Modeling:** Power Query (M) Excel
- **Visualization:** Excel (Advanced Charting)

Dataset Scope & Final Analytical Universe

This analysis is based on the TMDb 5000 Movie Dataset after full cleaning, validation, and relational restructuring. The table below defines the final analytical universe used across all phases.

ENTITY	RAW COUNT	FINAL VERIFIED COUNT
Movies	4,553	4,412 (Verified Theatrical)
Actors	~100k Unverified	100,261 (Linked to Financials)
Crew	~150k Unverified	59,214 (Big 5 Departments)

Only records with complete financial, relational, and role integrity were retained.

Phase 1: Metadata Architecture & Diagnostics

Initial inspection of **tmdb_5000_movies.csv** and **tmdb_5000_credits.csv** identified six critical structural failures hindering analysis.

Original Input	Corrected Name
Mercedes Mor\u00e1n	Mercedes Morán
Elizabeth Pe\u00f1a	Elizabeth Peña
Manny P\u00e9rez	Manny Pérez
Ciar\u00e1n Hinds	Ciarán Hinds
Daphn\u00e9e Duplaix Samuel	Daphnée Duplaix Samuel
Dagmara Domi\u0144czyk	Dagmara Domińczyk
Paul Rodr\u00edguez	Paul Rodríguez
Gin\u00e9s Garc\u00eda Mill\u00e1n	Ginés García Millán
Jean Michel Par\u00e9	Jean Michel Paré
Br\u00edan F. O'Byrne	Brían F. O'Byrne
P\u00e9ter Fancsikai	Péter Fancsikai
H\u00e9ctor Elizondo	Héctor Elizondo
Julio C. Pe\u00f1a	Julio C. Peña
Freddy Rodr\u00edguez	Freddy Rodríguez
Jeroen Krabb\u00e9	Jeroen Krabbé
Milo\u0161 Timotijevi\u0107	Miloš Timotijević
Alexander Skarsg\u00e5rd	Alexander Skarsgård

Data Architecture Summary

Methodology:

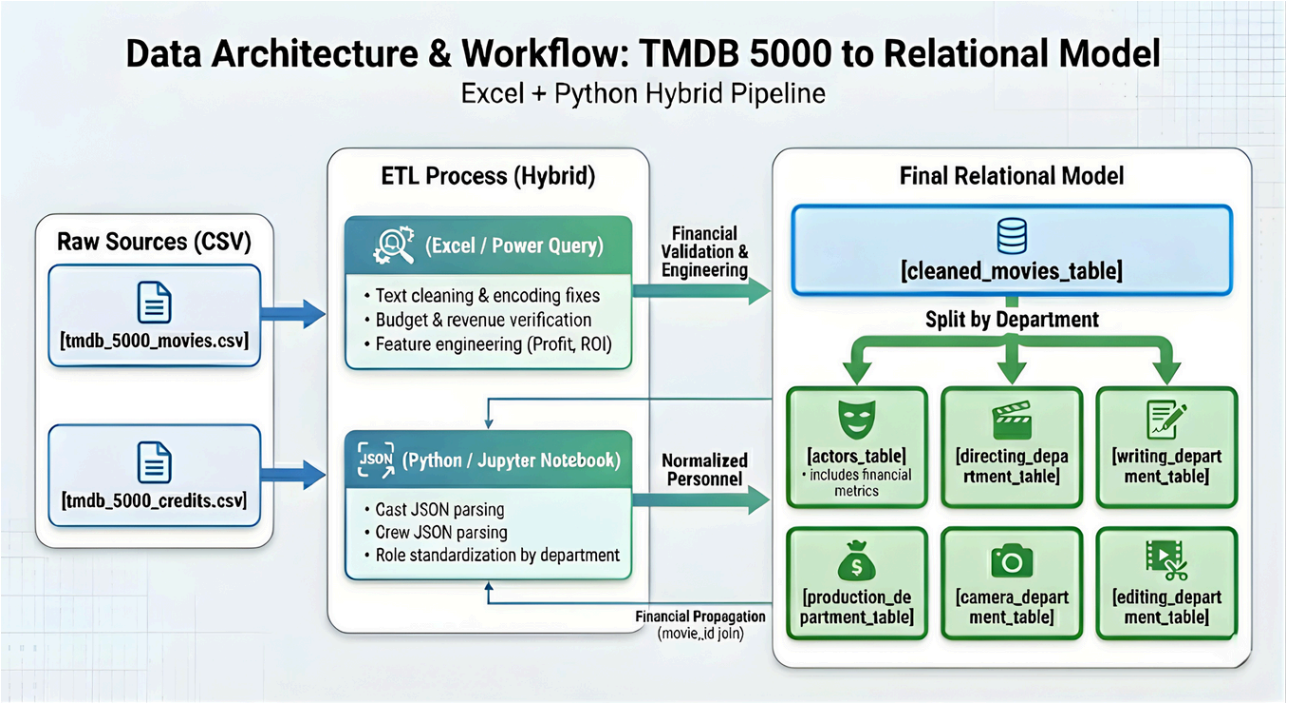
A Hybrid Excel–Python Pipeline.

- **Path A (Movies):**

  Processed in **Excel Power Query**. Focus on Metadata, Financial Verification, and Metric Engineering.

- **Path B (Credits):**

  Processed in **Python**. Focus on Iterative JSON Extraction and Relational Mapping.



✓ **STATUS: Architectural Baseline Established: Split-pipeline architecture validated.**

MODULE
Metadata Engineering

SOURCE
tmdb_5000_movies.csv

Implementation: Cleaning Movie Metadata

Step 1: Ingesting & Shaping (Power Query)

Logic: Reproducibility is critical. All transformations were recorded In the Power Query "Applied steps" layer to ensure the pipeline can be audited.

- **Trimming:** Recursive whitespace removal on all string fields.
- **Renaming Variables:** Standardized identifiers for Relational Integrity:
  - `id → movie_id(Primary Key)`
  - `original_title → title`
  - `budget → total_cost`

	movie_id	title	total_cost	revenue	genres
1	19995	Avatar	237000000	2787965087	Action
2	285	Pirates of the Caribbean: At World's End	300000000	961000000	Adventure
3	206647	Spectre	245000000	880674609	Action
4	49026	The Dark Knight Rises	250000000	1084939099	Action
5	49529	John Carter	260000000	284139100	Action
6	559	Spider-Man 3	258000000	890871626	Fantasy
7	38757	Tangled	260000000	591794936	Animation
8	99861	Avengers: Age of Ultron	280000000	1405403694	Action
9	767	Harry Potter and the Half-Blood Prince	250000000	933959197	Adventure
10	209112	Batman v Superman: Dawn of Justice	250000000	873260194	Action
11	1452	Superman Returns	270000000	391081192	Adventure
12	10764	Quantum of Solace	200000000	586090727	Adventure
13	58	Pirates of the Caribbean: Dead Man's Chest	200000000	1065659812	Adventure
14	57201	The Lone Ranger	255000000	89289910	Action
15	49521	Man of Steel	225000000	662845518	Action
16	2454	The Chronicles of Narnia: Prince Caspian	225000000	419651413	Adventure
17	24428	The Avengers	220000000	1519557910	Science Fiction
18	1865	Pirates of the Caribbean: On Stranger Tides	380000000	1045713802	Adventure
19	41154	Men in Black 3	225000000	624026776	Action
20	122917	The Hobbit: The Battle of the Five Armies	250000000	956019788	Action
21	1930	The Amazing Spider-Man	215000000	752215857	Action
22	20662	Robin Hood	200000000	310669540	Action
23	57158	The Hobbit: The Desolation of Smaug	250000000	958400000	Adventure

Step 2: Automated text Correction System

Risk: Leaving "Mojibake" (like: 'Ã¥') uncorrected would break the "Director Consistency" analysis by splitting a single director into two entities.

ENGINEERING LOGIC: The Translation Table

Applied a non-destructive translation layer using VLOOKUP to map corrupted strings to verified names:

=IFNA(VLOOKUP([@title], Corrections)_Table, 2, FALSE), [@title])

✓ **STATUS: Clean metadata standardized primary keys and encoded text.**

Financial Verification & Feature Engineering

Step 3: The “Preservation” Protocol

Business Risk: Deleting rows with $0.00 budgets would introduce “Survivorship Bias,” skewing the analysis toward big-budget films and hiding the efficiency of low-budget indie films.

> **VALIDATION CASE STUDY: Correction Mechanics**
>
> **Target:**
> *The Blair Witch Project* (Low Budget Outlier)
>
> - **Raw State:**
>   Budget = $0.00 (ROI Calculation Impossible)
> - **Enriched State:**
>   Budget = $60,000 (Source: The Numbers)
> - **Impact:**
>   Recovered a critical 4,000x ROI data points that defines the “Lottery Ticket” tier.

Step 4: Feature Engineering

Derived metrics were created to support the Executive Summary’s “Efficiency” findings.

- **Profit:** =[@revenue] - [@total_cost]
- **ROI:** =[@profit] / [@total_cost]
- **Standardization:** Decoded ISO languages and converted runtime integers (like: 90) to durations (1h 30m).

Outcome: Validated Movies Table

movie_id	title	language	revenue	total_cost	profit	return_of_investment	release_date	runtime	popularity	vote_average	vote
19995	Paranormal Activity	English	193355800	15000	193340800	12889.38667	39339	0.059722	47.456823	5.9	
285	The Blair Witch Project	English	248639099	60000	248579099	4142.984983	36355	0.05625	41.690578	6.3	
206647	Fifty Shades of Grey	English	571006128	300000	570706128	1902.35376	42046	0.086806	98.755657	5.2	
49026	Back to the Future Part II	English	332000000	300000	331700000	1105.666667	32832	0.075	43.345252	7.4	
49529	Lo imposible	English	180274123	200000	180074123	900.370615	41161	0.078472	47.559928	7	
559	Blade	English	131183530	170000	131013530	770.6678235	36028	0.083333	43.815492	6.5	
38757	A Bug's Life	English	363258859	500000	362758859	725.517718	36124	0.065972	87.350802	6.8	
9961	The Sixth Sense	English	672806292	1000000	671806292	671.806292	36378	0.074306	73.085576	7.7	
767	Sex Tape	English	126069509	250000	125819509	503.278036	41837	0.067361	72.641296	5.3	
209112	Pink Flamingos	English	6000000	12000	5988000	499	26370	0.064583	4.553644	6.2	
1452	Spy Kids 3-D: Game Over	English	197011982	400000	196611982	491.529955	37827	0.058333	21.998734	4.7	
10764	The Lake House	English	114830111	250000	114580111	458.320444	38884	0.06875	29.88736	6.5	
58	Super Size Me	English	28575078	65000	28510078	438.6165846	38003	0.069444	10.400603	6.6	
57201	The Gallows	English	42664410	100000	42564410	425.6441	42195	0.060417	18.045782	4.9	
49521	Open Water	English	54667954	130000	54537954	419.5227231	38205	0.054861	15.611857	5.4	
2454	The Brothers McMullen	English	10426506	25000	10401506	416.06024	34920	0.068056	1.578903	6.3	
24428	Shall We Dance?	English	170128460	420000	169708460	404.0677619	38275	0.074306	14.231899	5.9	
1865	The Other Woman	English	196781193	500000	196281193	392.562386	41745	0.075694	34.519636	6.2	
41154	Taken 2	English	376141306	1000000	375141306	375.141306	41179	0.063194	49.353524	6.1	
122917	The Texas Chain Saw Massacre	English	30859000	85000	30774000	362.0470588	27303	0.057639	29.262427	7.2	
1930	The Firm	English	270248367	750000	269498367	359.331156	34150	0.106944	23.736535	6.6	
20662	He's Just Not That Into You	English	177259441	500000	176759441	353.518882	39850	0.089583	26.253357	6.2	
57158	Gothika	English	141591324	400000	141191324	352.97831	37946	0.068056	27.303608	5.8	
2268	Mr. Deeds	English	171269535	500000	170769535	341.53907	37435	0.066667	18.121404	5.6	
254	The Flintstones	English	341631208	1000000	340631208	340.631208	34480	0.063194	28.964162	5	
597	Bridge of Spies	English	165478348	500000	164978348	329.956696	42292	0.097917	48.445978	7.2	
271110	Unknown	English	130786397	400000	130386397	325.9659925	40590	0.078472	38.643914	6.5	
44833	Taken 3	English	325771424	1000000	324771424	324.771424	41989	0.075694	88.844777	6.1	
135397	The Conjuring 2	English	320170008	1000000	319170008	319.170008	42503	0.093056	68.794673	7	
37724	Flightplan	English	223387299	700000	222687299	318.1247129	38617	0.068056	32.05145	6.1	
558	Sin City	English	158733820	500000	158233820	316.46764	38443	0.086111	66.003433	7.2	
68721	The Proposal	English	317375031	1000000	316375031	316.375031	39965	0.075	36.238968	6.7	
12155	Indiana Jones and the Last Crusade	English	474171806	1500000	472671806	315.1145373	32652	0.088194	80.972475	7.6	
36668	Bambi	English	267447150	858000	266589150	310.709965	15567	0.048611	47.651878	6.8	
62211	Big Momma's House 2	English	138259062	450000	137809062	306.24236	38743	0.06875	17.61292	5.4	

✓ **STATUS: Final Output: 4,412 Verified Records with 100% Financial Density.**

PHASE
02. JSON Parsing

STACK
PYTHON JSON

Architectural Migration: The Python Shift

> **Phase Scope:** Auditing the parsing of 100,000+ personnel records.
>
> **Primary Risk:** Loss of referential integrity due to orphan records.

While Excel handled metadata, the credits.csv file presented as structural bottleneck: **Nested JSON Arrays.** Attempting to parse 100,000+ crew members within Excel cells caused instability.

> **STRATEGIC DECISION: Migration to Jupyter Notebook**
>
> To ensure pipeline robustness, I migrated the parsing logic to Python. This allowed for iterative extraction of nested objects without risking data corruption.

Module 1: The Cast Pipeline

I developed a custom iterator parse_cast to flatten the JSON structure. The algorithm extracts actor names and maps them to the movie_id foreign key.

ALGORITHMIC LOGIC SUMMARY

```
01. Iterate through dataframe rows.
02. Deserialize JSON string: json.loads(row['cast']).
03. Extract attributes: [name, id, character].
04. Append foreign key: Link entity to current movie_id.
```

[8]:
```
actor_rows = []

for _, row in credits.iterrows():
    actor_rows.extend(parse_cast(row["cast"], row["movie_id"]))

actors_df = pd.DataFrame(actor_rows)
actors_df.head()
```

[8]:

	movie_id	actor_name
0	19995	Sam Worthington
1	19995	Zoe Saldana
2	19995	Sigourney Weaver
3	19995	Stephen Lang
4	19995	Michelle Rodriguez

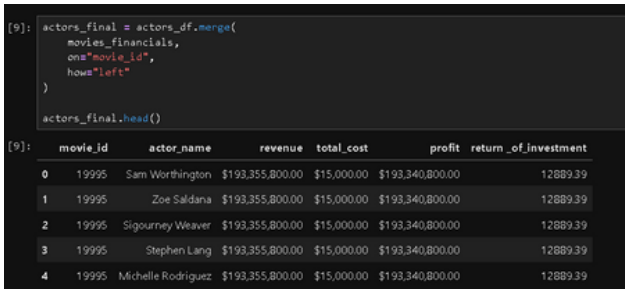
✓ **STATUS: Outcome: 100,261 Actor records normalized.**

Referential Integrity & Crew Logic

Critical Logic: Cascading Deletions

A major risk in relational modeling is the “Orphan Record”. Since I deleted 141 movies in Phase 1, any actors linked *only* to those movies had to be removed.

> **ENGINEERING ACTION: Enforcing Consistency**
>
> I applied a “Left Join” against the verified movie list. Any personnel record resulting in a null financial value was purged.
>
> Associated Risk: Orphan records causing invalid ROI attribution.



```
[9]: actors_final = actors_df.merge(
         movies_financials,
         on="movie_id",
         how="left"
     )

     actors_final.head()
```

[9]:

	movie_id	actor_name	revenue	total_cost	profit	return_of_investment
0	19995	Sam Worthington	$193,355,800.00	$15,000.00	$193,340,800.00	12889.39
1	19995	Zoe Saldana	$193,355,800.00	$15,000.00	$193,340,800.00	12889.39
2	19995	Sigourney Weaver	$193,355,800.00	$15,000.00	$193,340,800.00	12889.39
3	19995	Stephen Lang	$193,355,800.00	$15,000.00	$193,340,800.00	12889.39
4	19995	Michelle Rodriguez	$193,355,800.00	$15,000.00	$193,340,800.00	12889.39

movie_id	actor_name	revenue	total_cost	profit	return_of_investment
7978	Benicio del Toro				
7978	Anthony Hopkins				
7978	Emily Blunt				
7978	Hugo Weaving				
7978	Cristina Contes				
7978	David Sterne				
7978	Elizabeth Croft				
7978	Simon Merrells				
7978	Art Malik				
7978	Asa Butterfield				
7978	Geraldine Chaplin				
7978	Olga Fedori				
7978	Lorraine Hilton				
7978	John Owens				
10357	Tommy Lee Jones				
10357	Anne Heche				
10357	Gaby Hoffmann				
10357	Don Cheadle				
10357	Joshua Fardon				
10357	Jacqueline Kim				
10357	Keith David				
10357	John Corbett				

Module 2: The Crew Pipeline (Signal-to-Noise)

The raw crew data contained thousands of roles irrelevant to financial strategy. To focus the analysis on “Creative ROI,” I implemented a Departmental Filter.

The “Big 5” Department Strategy

I filtered the master dataset to isolate the five departments with the highest potential impact on box office performance.

- **Directing** (Creative Leadership)
- **Writing** (Story Quality)
- **Production** (Logistics & Budgeting)
- **Camera** (Visual Quality)
- **Editing** (Pacing & Final Cut)

[17]:
```
directing_df = crew_master[
    crew_master["department"] == "Directing"
]

directing_df.to_csv("directing_department_table.csv", index=False)
```

[18]:
```
writing_df = crew_master[
    crew_master["department"] == "Writing"
]

writing_df.to_csv("writing_department_table.csv", index=False)
```

[19]:
```
production_df = crew_master[
    crew_master["department"] == "Production"
]

production_df.to_csv("production_department_table.csv", index=False)
```

[20]:
```
camera_df = crew_master[
    crew_master["department"] == "Camera"
]

camera_df.to_csv("camera_department_table.csv", index=False)
```

[21]:
```
editing_df = crew_master[
    crew_master["department"] == "Editing"
]

editing_df.to_csv("editing_department_table.csv", index=False)
```

✓ **STATUS: Outcome: Five department-specific crew tables created.**

Quality Assurance & Final Schema

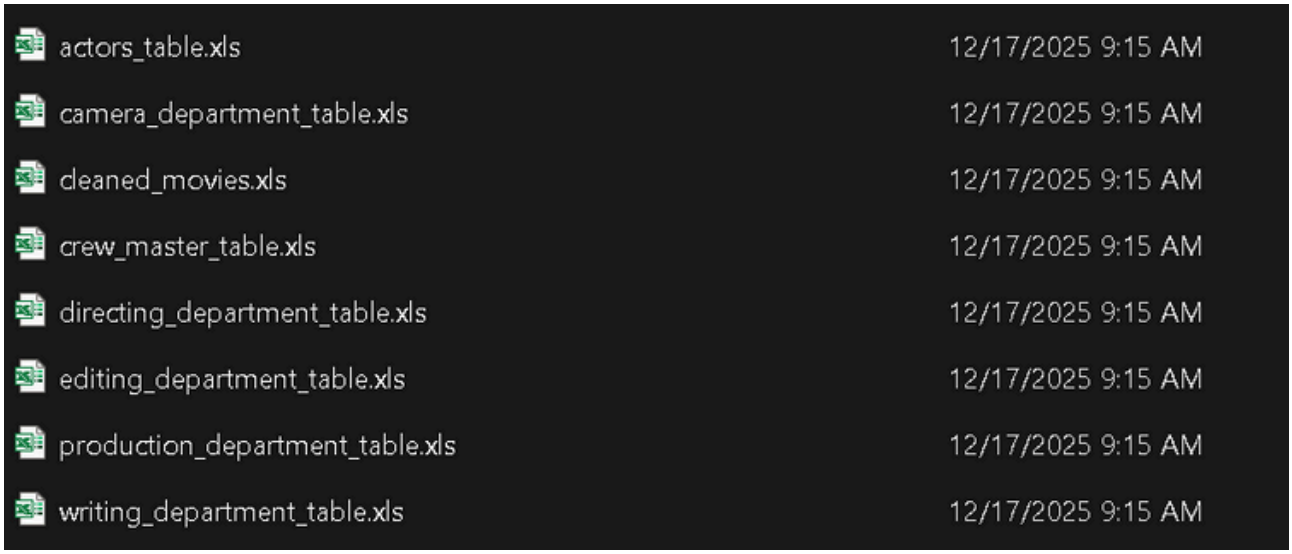
The final step involved re-importing the Python-generated CSVs into Excel for a final redundancy check. I utilized the **Remove Duplicates** tool, configuring it to check strict equality across all columns.

Unified QA Matrix: Redundancy & Integrity

Table Entity	Cascading Deletions (Orphans)	Duplicate Rows Removed	Final Verified Records
Actors	-5,829	-167	100,261
Directing	-528	-5	7,616
Writing	-654	-628	9,405
Production	-1,278	-318	26,079
Camera	-377	-160	8,667
Editing	-353	-7	7,495

Final Output: The Relational Ecosystem

The pipeline successfully transformed two flat CSV files into a star-schema style relational model.



EXPLICIT ANALYTICAL EXCLUSIONS

To preserve internal consistency, the following elements were intentionally excluded:

- **Streaming/Ancillary Revenue:**
  Analysis is strictly Box Office to avoid speculative data.
- **Inflation Adjustment:**
  All values are Nominal USD (Historic) to reflect capital decisions at time of greenlight.
- **Granular Marketing Spend:**
  Excluded due to incomplete reporting in the source dataset.

PHASE
03. Strategic Intelligence

FOCUS
Macro-Economics

> **METRIC STABILITY & ROBUSTNESS**
> All core rankings and derived metrics were recomputed after data cleaning and orphan removal. Rank ordering remained materially stable, indicating that observed patterns are driven by underlying data structure rather than preprocessing artifacts.

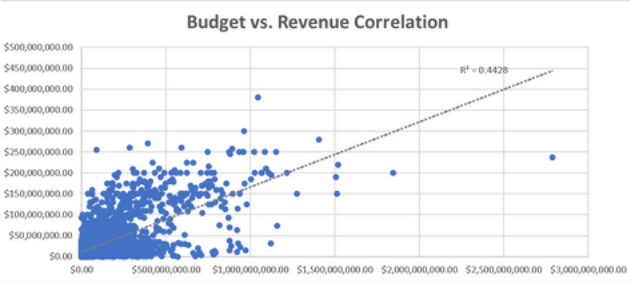
Macro-Economics: The Efficiency Paradox

The data reveals a distinct inverse relationship between capital deployment and percentage return.

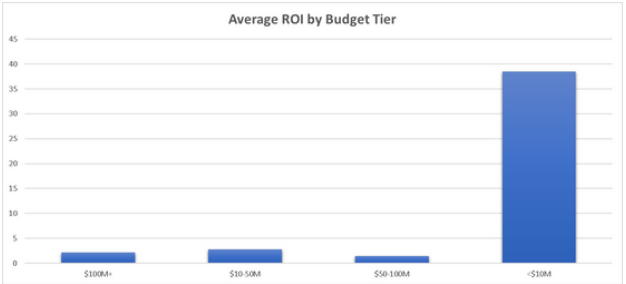
Insight 1: The "Efficiency vs. Scale" Trade-off

Low Budget (<$10 Million): Highest Efficiency Tier. Low risk with highest percentage ceiling.
Blockbusters ($100 Million Plus): Low relative efficiency. These are "high maintenance" assets that require massive marketing to break even.

> *"Executive Translation: You generally have to 'spend money to make money' (Revenue), but as you scale, your margin for error disappears."*



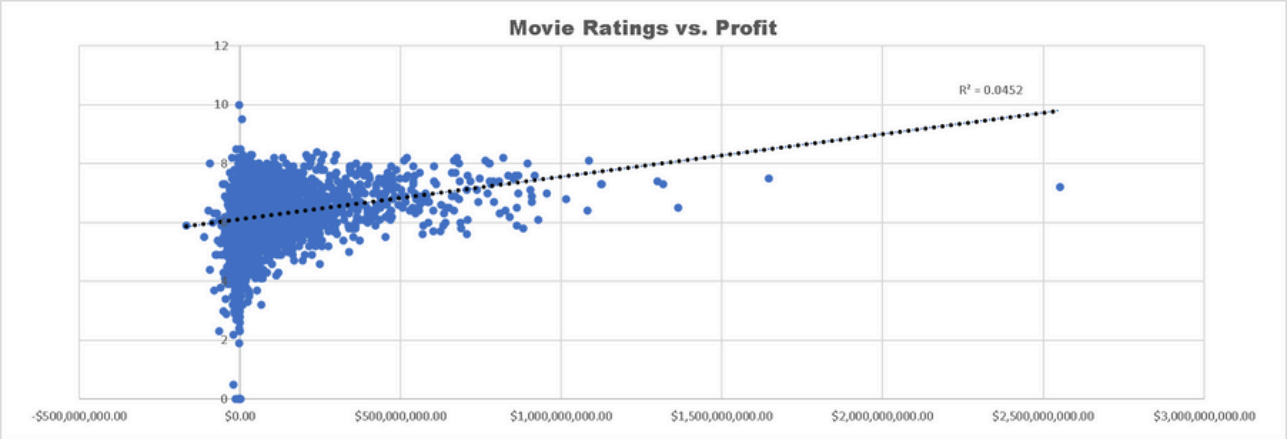
Analysis indicates a moderate correlation for Revenue



ROI Cliff: Efficiency drops as costs rise.

Contrarian Insight: The “Quality” Myth

Regression analysis reveals a weak relationship between critical reception (User Ratings) and financial success.

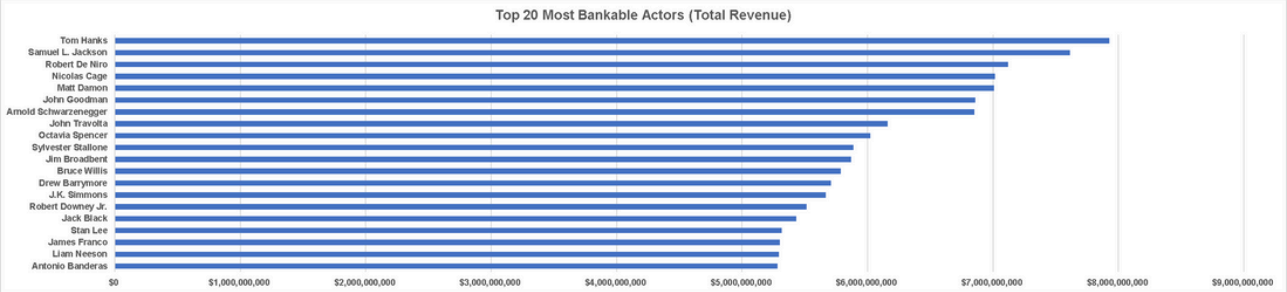


Talent Intelligence: The Three Pillars

I structured the talent analysis around three distinct financial KPIs: **Volume**, **Efficiency**, and **Risk.**

1. Actor Bankability (Cash Flow)

Goal: Identify "Star Power"—who guarantees opening weekend volume?

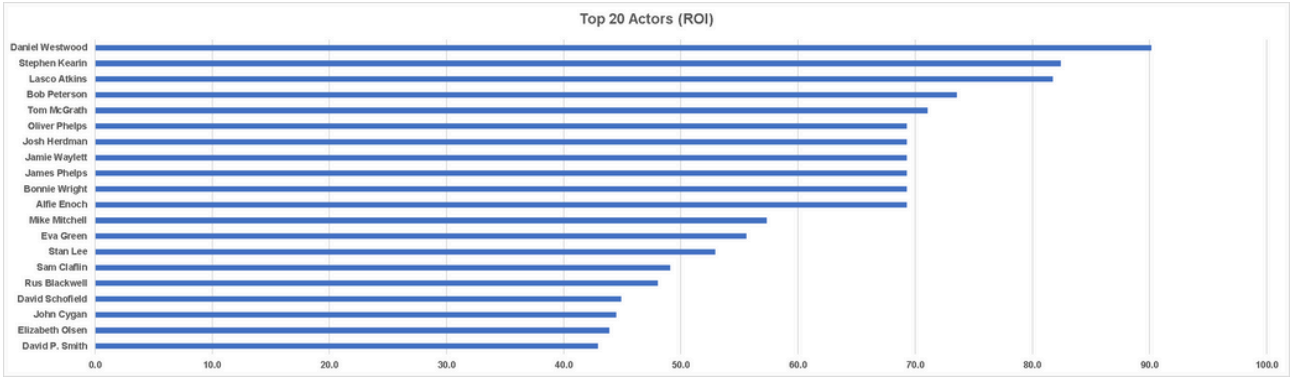


Leader: Tom Hanks. Defines "Lead Star Efficiency".

2. Portfolio ROI (Capital Efficiency)

Goal: Identify "Hidden Gems"—who maximizes every dollar of budget?

```
Metric Formula: Aggregate ROI = ((Sum/Revenue) - Sum(Cost)) / Sum(Cost)
```



ARCHETYPE	ACTOR	METRIC
Efficiency King	Daniel Westwood	**High ROI Multiplier** (Low Cost / High Return)
Volume Star	Tom Hanks	**Standard ROI Multiplier** (High Cost / High Return)

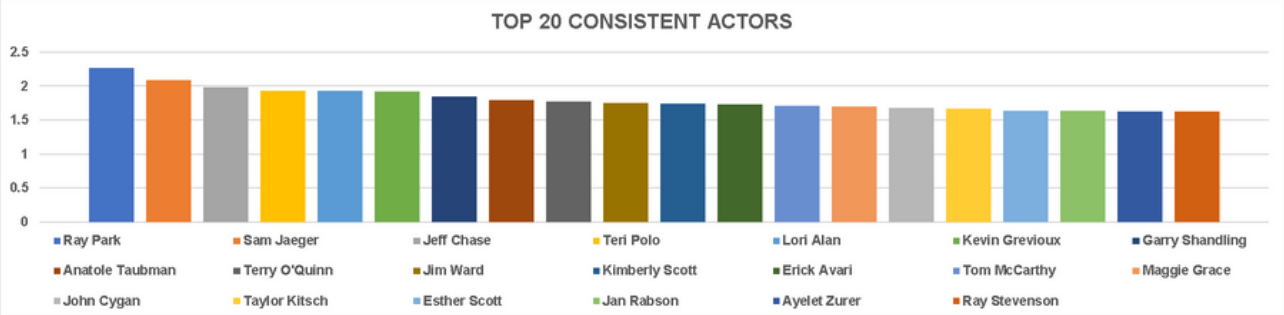
3. Consistency Scoring (Risk Mitigation)

Goal: Who is 'Safe'? Low volatility assets.

Methodology: The Consistency Ratio

```
Score = Average Profit / Standard Deviation of Profit
```

Observation: Ray Park exemplifies the low-volatility archetype. Niche franchise actors offer lower financial volatility.

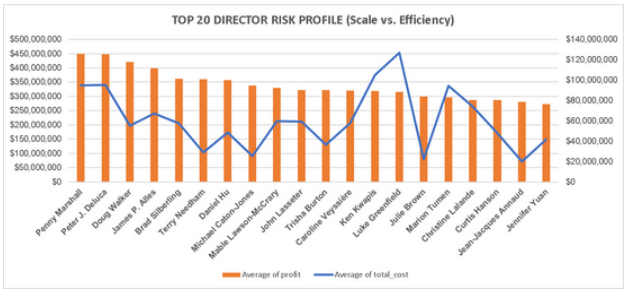


✓ **STATUS:** Outcome: "Three-Pillar" Talent Valuation Model established.

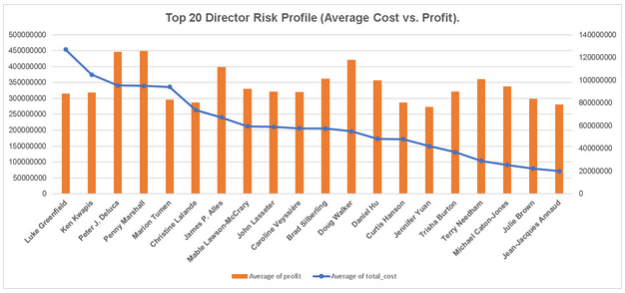
Creative Leadership & Risk Profiling

Director Risk Profiles

Do bigger budgets yield bigger profits? The “Risk Profile” chart (Cost Line vs. Profit Bar) expose the truth.



Efficiency Masters: Terry Needham.



Money Pits: High capital outlay with diminishing returns.

Writer Analysis: The “Golden Pen”

Insight: The highest cumulative profits belong to creators of Source Material (IP), not screenwriters.

- **#1 Stan Lee:** Ownership of the Marvel IP yields the highest financial ceiling.
- **#2 Steve Kloves:** Retaining a single voice (Harry Potter) reduces volatility.

Row Labels	Sum of profit
Stan Lee	
Sylvester Stallone	
Jack Kirby	
Akiva Goldsman	
Michael France	
Eric Roth	
Brian Helgeland	
Steven Zaillian	
David Koepp	
Bob Kane	
Zak Penn	
Mark Bomback	
Lilly Wachowski	
Lana Wachowski	
Timothy Burgard	
Joss Whedon	
John Lasseter	
Robert Zemeckis	
Scott Frank	
Stephen Gaghan	

Row Labels	Average of profi	StdDev of profi	Consistency Score
John Rogers	$ 63,794,794	$ 23,250,976	2.74
Steve Kloves	$ 203,468,701	$ 76,493,415	2.66
Craig Mazin	$ 116,987,408	$ 44,922,936	2.60
Jonathan Lemki	$ 118,903,814	$ 46,613,576	2.55
Brad Bird	$ 114,231,965	$ 45,895,330	2.49
Evan Daugherty	$ 146,605,326	$ 59,705,704	2.46
Barry W. Blaust	$ 57,133,280	$ 26,161,344	2.18
Steve Ditko	$ 214,228,348	$ 99,634,743	2.15
Chris Henchy	$ 117,878,627	$ 54,880,718	2.15
Alfred Gough	$ 129,337,049	$ 61,797,734	2.09
Miles Millar	$ 129,337,049	$ 61,797,734	2.09
John Collee	$ 80,977,958	$ 42,943,394	1.89
Paul Weitz	$ 56,546,252	$ 31,072,150	1.82
Jonathan Hensl	$ 43,144,004	$ 23,799,443	1.81
William Broyles	$ 183,162,961	$ 102,488,203	1.79
William Monaha	$ 101,834,808	$ 57,076,300	1.78
J.R.R. Tolkien	$ 90,739,137	$ 51,695,689	1.76
Nicholas Stoller	$ 89,390,508	$ 51,479,445	1.74
James L. Brook	$ 229,427,222	$ 132,573,643	1.73
Jason Segel	$ 109,355,055	$ 63,465,420	1.72

✓ **STATUS:** Outcome: Risk Profiles quantified for Directors and Writers.

Operational Strategy & Conclusion

The "Technical Alpha" (Crew Analysis)

My analysis of the "Big 5" departments confirms that **Consistency Drives Profit.**

Role	Key Insight	Top Performer
Producers	Animation/Toys drive highest margins.	Brian Goldner (Hasbro)
Camera	Studios retain visual teams for brand continuity.	Jay Maidment (Marvel)
Editors	"Architects of Pacing"—critical for franchise retention.	Jabez Olssen

Strategic Recommendations for Capital Allocation

1. The "Portfolio" Strategy: Mix "High-Budget/High-Revenue" anchors with "Low-Budget/High-ROI" bets.
(Derived from Phase 3: Budget Elasticity Analysis)

2. Prioritize IP over Scripts: "Original Writers" generate higher Lifetime Value than hired screenwriters.
(Derived from Phase 3: Writer Profitability Matrix)

3. Marketing > Perfection: Investment should be weighted toward distribution reach rather than artistic reshoots.
(Derived from Phase 3: Ratings vs. Profit Regression)

Final Conclusion

This project demonstrates the full value chain of data analytics: from the rigorous engineering of 4,553 raw records to the derivation of high-level capital strategies. Raw data is noise; engineered data is a decision-support system.

✓ **STATUS:** Master Audit Complete. Decision Framework Ready.

SUPPORTING DOCUMENTATION

Full Technical Walkthrough

The following section contains the raw engineering logs, extended code snippets, and intermediate transformation screenshots supporting the audit findings.

APPENDIX

Navigate to: Detailed Step-by-Step Execution