# Documentation & Data Dictionary: IMDb and Box Office Mojo

IMDb Developer



2020-05-01

**Contents**

[Documentation & Data Dictionary: IMDb and Box Office Mojo](#_bookmark0) 4

[Getting Support](#_bookmark1) 5

[Key Concepts](#_bookmark2) 5

[IMDb IDs](#_bookmark3) 5

[JSON Lines File Format](#_bookmark4) 5

[Versioning](#_bookmark5) 5

[Data Structure Conventions](#_bookmark6) 6

[Changes to Entities and Resolving IDs](#_bookmark7) 6

[Data Consistency Model](#_bookmark8) 7

[Linking to IMDb](#_bookmark9) 7

[Data Dictionary - Name Essential](#_bookmark10) 8

[nameId](#_bookmark11) 8

[name](#_bookmark12) 8

[awards](#_bookmark13) 8

[filmography](#_bookmark14) 8

[knownFor](#_bookmark15) 10

[Data Dictionary - Title Essential](#_bookmark16) 11

[titleId](#_bookmark17) 11

[remappedTo](#_bookmark18) 11

[originalTitle](#_bookmark19) 11

[akas](#_bookmark20) 11

[awards](#_bookmark21) 11

[creditsByCategory](#_bookmark22) 12

[principalCastMembers](#_bookmark23) 14

[principalCrewMembers](#_bookmark24) 14

[certificates](#_bookmark25) 15

[companies](#_bookmark26) 15

[countries](#_bookmark27) 16

[episodeInfo](#_bookmark28) 16

[seriesInfo](#_bookmark29) 16

[episodeTitleIds](#_bookmark30) 17

[officialSiteLinks](#_bookmark31) 17

[genres](#_bookmark32) 17

[image](#_bookmark33) 18

[imdbUrl](#_bookmark34) 18

[isAdult](#_bookmark35) 18

[keywords](#_bookmark36) 18

[languages](#_bookmark37) 18

[locations](#_bookmark38) 19

[movieConnections](#_bookmark39) 19

[plot](#_bookmark40) 20

[plotShort](#_bookmark41) 20

[plotMedium](#_bookmark42) 20

[plotLong](#_bookmark43) 20

[releaseDates](#_bookmark44) 20

[Production Status](#_bookmark45) 21

[runtimeMinutes](#_bookmark46) 21

[taglines](#_bookmark47) 22

[titleType](#_bookmark48) 22

[imdbRating](#_bookmark49) 22

[year](#_bookmark50) 22

[Data Dictionary - Box Office](#_bookmark51) 23

[boxoffice\_title\_budgets\_v1](#_bookmark52) 23

[boxoffice\_title\_grosses\_v1](#_bookmark53) 23

[boxoffice\_title\_opening\_weekends\_v1](#_bookmark54) 23

[boxoffice\_title\_grosses\_timeseries\_v1](#_bookmark55) 24

[Querying Data in Amazon Athena Tables](#_bookmark56) 26

[What is Amazon Athena?](#_bookmark57) 26

[Getting Started](#_bookmark58) 26

[What Are the Highest-Rated Movies On IMDb?](#_bookmark59) 26

[What Are the Title Texts for the Titles That a Person Is Known For?](#_bookmark60) 27

[What Are the Names of the Principal Cast for a Title?](#_bookmark61) 28

[What Awards was a Title Nominated for, and who were the Award Nominees?](#_bookmark62) 29

[What Are the Title Texts for Episodes of a Series?](#_bookmark63) 30

[What Is the Title Text of the Series for an Episode Title?](#_bookmark64) 31

[Appendix 1: Box Office Mojo Areas](#_bookmark65) 32

[Area Rollups and Special Areas](#_bookmark66) 32

[Individual Areas](#_bookmark67) 33

**Documentation & Data Dictionary: IMDb and Box Office Mojo**

Access IMDb’s metadata for every movie, TV series and video game title as well as performers and creators, along with full lifetime box office grosses from IMDbPro’s Box Office Mojo. Our bulk data access products help entertainment fans share their passion with the world, including IMDb’s 1-10 star rating, a daily-computed average of votes from IMDb’s global audience of over 250 million fans.

Our full range of bulk data products are delivered through [Amazon Data Exchange](https://aws.amazon.com/marketplace/search/results?page=1&filters=FulfillmentOptionType%2Cvendor_id&FulfillmentOptionType=AWSDataExchange&vendor_id=0af153a3-339f-48c2-8b42-3b9fa26d3367) and updated daily. This document explains schema and documentation for all IMDb content available in bulk via Amazon Data Exchange. Data types are packaged across several different data products ranging from IMDb’s world-renowned 1-10 star rating, to full industry Box Office grosses from IMDbPro’s Box Office Mojo. If you have any questions on our documentation, or would like to access a free sample of our data products, please get in touch via [developer.imdb.com](https://developer.imdb.com/).

## Getting Support

For technical support or licensing questions please email [imdb-licensing@imdb.com.](mailto:imdb-licensing@imdb.com)

## Key Concepts

### IMDb IDs

IMDb uses unique identifiers for each of the entities referenced in IMDb data. For example we have “Name IDs” identifying name entities (people) and “Title IDs” identifying title entities (movies, series, episodes and video games). IMDb’s identifiers always take the form of two letters, which signify the type of entity being identified, followed by a sequence of at least seven numbers that uniquely identify a specific entity of that type. For example:

* tt0050083 is the unique identifier for the movie “12 Angry Men (1957)”, where tt signifies that it’s a title entity and 0050083 uniquely indicates “12 Angry Men (1957)”.
* nm0000020 is the unique identifier for the actor “Henry Fonda”, where nm signifies that it’s a name entity and 0000020 uniquely indicates “Henry Fonda”.

Within the data set, each entry relates to a single IMDb identifier.

### JSON Lines File Format

IMDb’s data set is provided in JSON Lines file format. The files are UTF-8 encoded text files, where each line in the file is a valid JSON string. Each JSON document, one per line, relates to a single entity, uniquely identified by an IMDb ID. We also provide a JSON schema that documents the format that is used for each JSON document within the file.

### Versioning

Every published revision of IMDb’s data set contains data file(s), documentation for that data, and a schema which validates that data. Each of these is associated with a correlated version number, which can be found at the end of their filenames.

At any time we may change the format of new data set revisions and their accompanying schema, but previously published data set revisions will remain unchanged. If data from a new revision of the data set is not compatible with the previous schema (i.e. a breaking change) then we will increment the version number for the data files, schema, and documentation. In this case we will publish both

formats of the data set for some period of time before we stop publishing the older one. The data set format and schema may change *without* incrementing the schema version number if the change is compatible with previous revisions (i.e. a non-breaking change).

The following are examples of *non-breaking* changes to the schema:

* Adding a new key anywhere in the structure.
* Removing an optional key.
* Changing a key from optional to required.
* Changing the validation rules for a specific key such that all values still validate against previous validation rules.

The following are examples of *breaking* changes to the schema:

* Changing a key from required to optional.
* Removing a required key.
* Changing the validation rules for a specific key such that newly published values may exist that do not validate against previous validation rules.

### Data Structure Conventions

There are some conventions you should be aware of when using IMDb’s data set:

* There are no null values in the data set. If we do not have a value for a particular key we omit publishing that key. Keys which are required by the schema will never have a null value.
* There are no empty objects in the data set. If an object would have contained no keys we omit publishing that object.
* There are no empty arrays in the data set. If an array value would have contained no items we omit publishing the corresponding key.

### Changes to Entities and Resolving IDs

#### Duplicate IDs

IMDb’s data set is constantly being updated, adding more data and improving the quality of the data we have. While there is only ever one entry per IMDb ID, we sometimes find that we have duplicate IMDb IDs for an entity within our system. For example, we may learn that two people we have iden- tified separately are actually the same person. When this happens, we maintain the data associated

with both identifiers in the data set, duplicating the data. This allows you to continue using any match- ing you have between IMDb identifiers and other identifiers. To identify when this is the case we in- clude a remappedTo field on one of the copies which gives you the new preferred identifier for that entity.

#### Deleted IDs

Sometimes we delete entities from the data set. The most prominent example of this is the deletion of titles that have been canceled during development and will therefore never be released. When we delete an entity it is no longer included in the data set. The identifier associated with it is never reused for a different entity.

### Data Consistency Model

IMDb’s data is constantly being expanded and updated, and it can take seconds or minutes for a change to propagate throughout the entire catalog. This means that the snapshot of IMDb’s data published may contain temporary inconsistencies. For example, it is possible that we report an ac- tor appearing in a title in their filmography, but it has not yet propagated to that title’s credits. Each individual inconsistency will be resolved in the next published revision of the data set.

### Linking to IMDb

IMDb’s data contains URLs that you can use to link back to the IMDb website in any experience you build for your users. Your license may require you to attach a “refmarker” to the end of the URL. The “refmarker” is a special sequence of characters that we use to identify the source of our traffic. Add the “refmarker” to the URL by appending ?ref\_=xx\_xxx\_x to the URL, where xx\_xxx\_x is replaced by the code we have provided to you. A full URL could look something like [https://www.imdb.com/title/tt0050083/?ref\_=my\_ref\_marker](http://www.imdb.com/title/tt0050083/?ref_=my_ref_marker).

## Data Dictionary - Name Essential

### nameId

The unique IMDb ID for the name in question. Each IMDb ID appears exactly once.

### name

The primary name by which this person is known, usually the one by which they are most often cred- ited. For more information about how IMDb defines the primary name see [IMDb’s help site](https://help.imdb.com/article/contribution/names-biographical-data/names/GSA3M6SFHRAERXZ3).

### awards

A list of awards that this person has won or been nominated for. This includes the name and category of the award, the name and year of the award event, and whether the person won the award. Note that ‘winner’ may be false because the person is known not to have won the award (where the awards event occured in the past) or because the winner is not yet known (where the awards event occurs in the future, but the nominations have been announced). If ‘winner’ is true that means the awards event has already occured, and the person won the award.

#### Example

{

"awards": [

{

"year": 1958,

"awardName": "Golden Globe", "category": "Best Actor - Drama", "winner": **false**,

"event": "Golden Globes, USA"

},

**...**

]

}

### filmography

The filmography for this name as a list of credits. Each credit is within a “category” such as “ac- tress”, “director” or “editorial\_department”. For cast categories (e.g. “actor”), we include the roles

that the person played and the billing they had in the end credits (if available). For crew categories (e.g. “writer”) we include the more specific “jobs” that the person was credited with if applicable. Lists of credits, roles, and jobs are each in on-screen credits order.

Credits can have a list of attributes. At the moment we provide the following attributes:

* “uncredited”: signals that while the person performed this role on the title, they were not present in the title’s end credits.
* “voice”: signals that this person provided a voice only performance for this title.

Additional information about these attributes can be found on [IMDb’s help site](https://help.imdb.com/article/contribution/filmography-credits/cast-acting-credits-guidelines/GH3JZC74FVYKKFMD).

For episodic credits we only include the series in the list. To get full information on which episodes were worked on, look at the episode credits in the title file.

#### Example

A cast credit

{

"filmography": [

{

"titleId": "tt0050083", "category": "actor", "billing": 8,

"roles": ["Juror 8"]

},

**...**

]

}

A crew credit

{

"filmography": [

{

"titleId": "tt0052462", "category": "producer", "jobs": ["executive producer"]

},

**...**

]

}

An uncredited credit

{

"filmography": [

{

"titleId": "tt0050083", "category": "actor",

"roles": ["Judge"], "attributes": ["uncredited"]

}

**...**

]

}

### knownFor

A short list of IMDb title IDs for the titles in which this person is most well known for being involved, and the category of job that they had on that title (e.g. “actor” or “director”). This is always a subset of filmography but the selection and order is determined by IMDb. For more details see [IMDb’s help](https://help.imdb.com/article/imdb/discover-watch/known-for-title-selection/GNL2E4LJVKM9QLMD) [site](https://help.imdb.com/article/imdb/discover-watch/known-for-title-selection/GNL2E4LJVKM9QLMD). For further details on their involvement see the filmography entry, or the creditsByCat- egory entry on the title in question.

#### Example

{

"knownFor": [

{

"titleId": "tt0050083", "category": "actor"

},

**...**

]

}

## Data Dictionary - Title Essential

### titleId

The unique IMDb ID for the title in question. Each IMDb ID appears exactly once.

### remappedTo

If there are duplicate title entities for this title, the IMDb ID of the primary title entity for this title. See “Duplicate IDs” in the “Key Concepts” section of this document for more information.

### originalTitle

The original title text of the title, normally what the title is known as in its original country of release.

### akas

A list of alternative title texts by which this title is also known. Each title is listed with additional in- formation about the usage of that title text, e.g. what region it is from, and what language it is used in.

#### Example

{

"akas": [

{

"title": "12 Homens e uma Sentença", "region": "BR"

},

**...**

]

}

### awards

A list of awards that this title has won or been nominated for. This includes the name and category of the award, the name and year of the award event, and whether the title won the award. Note that

‘winner’ may be false because the title is known not to have won the award (where the awards event occured in the past) or because the winner is not yet known (where the awards event occurs in the future, but the nominations have been announced). If ‘winner’ is true that means the awards event has already occured, and the title won the award.

#### Example

{

"awards": [

{

"awardName": "Oscar", "category": "Best Picture", "event": "Academy Awards, USA", "winner": **false**,

"year": 1958

},

**...**

]

}

### creditsByCategory

The credits for this title organized by category. Each entry in this list represents a single category and gives you a list of credits within that category. For cast credits we include the roles that the person played and the billing they had in the end credits (if available). For crew credits we include the more specific “jobs” that the person was credited with if applicable. Lists of credits, roles, and jobs are each in on-screen credits order.

Credits can have a list of attributes. At the moment we provide the following attributes:

* “uncredited”: signals that while the person performed this role on the title, they were not present in the title’s end credits.
* “voice”: signals that this person provided a voice only performance for this title.

Additional information about these attributes can be found on [IMDb’s help site](https://help.imdb.com/article/contribution/filmography-credits/cast-acting-credits-guidelines/GH3JZC74FVYKKFMD).

For series we include anyone who is credited on any episode. To get full information on which episodes were worked on by a specific person, look at the episode credits.

#### Example

A crew category

{

"creditsByCategory": [

{

"category": "sound\_department", "credits": [

{

"nameId": "nm0322302", "jobs": ["sound"]

},

{

"nameId": "nm0334505",

"jobs": ["re-recording mixer"], "attributes": ["uncredited"]

}

]

},

**...**

]

}

A cast category

{

"creditsByCategory": [

{

"category": "cast", "credits": [

{

"nameId": "nm0000842", "roles": ["Juror 1"],

"billing": 1

},

{

"nameId": "nm0275835", "roles": ["Juror 2"],

"billing": 2

},

**...**

]

},

**...**

]

}

### principalCastMembers

A short list of the most important cast credits for this title. This is always a subset of the cast from the creditsByCategory list, but the selection and order is determined by IMDb. Often it is similar to top-billed cast but it can be different, for example if the title credits are in order of appearance or alphabetical. For more details see [IMDb’s help site](https://help.imdb.com/article/contribution/filmography-credits/how-are-cast-credits-ordered-why-don-t-the-main-stars-appear-at-the-top-of-the-cast/G39K5N4YYV2QJ4GR). Also includes the role or roles played (in on-screen credits order) and the billing in the full cast list.

#### Example

{

"principalCastMembers": [

{

"nameId": "nm0000020", "category": "actor",

"roles": ["Juror 8"],

"billing": 8

},

**...**

]

}

### principalCrewMembers

A short list of the most important crew credits for this title. This is always a subset of the crew from the creditsByCategory list, but the selection and order is determined by IMDb. Also includes the category and job which qualified the credit for this list.

#### Example

{

"principalCrewMembers": [

{

"nameId": "nm0741627", "category": "writer",

"job": "story"

},

**...**

]

}

### certificates

A list of content rating certifications that have been given to a title, and the region where the rating applies or applied. For example a title may be given a ‘PG-13’ rating in the ‘US’ region (by the MPAA). There may be additional attributes about the certificate or reasons for the rating provided by the rating organization.

#### Example

{

"certificates": [

{

"region": "US",

"rating": "TV-PG"

},

**...**

]

}

### companies

Lists of the names of distribution, production, special-effects, and other miscellaneous companies as- sociated with the making or subsequent distribution of this title. This list includes all companies that have ever been involved with the title, even if their involvement has now ended. Additional informa- tion about companies associated with titles can be found on [IMDb’s help site](https://help.imdb.com/article/contribution/industry-professional/companies-data/GM45NQ429VURNCSG). These are ordered by on-screen credit order, or in the case of distribution companies by distribution release date.

#### Example

{

"companies": {

"distribution": ["United Artists", "CBS/Fox", "Warner Home Video"], "production": ["Orion-Nova Productions"],

"miscellaneous": [

"International Alliance of Theatrical Stage Employees (IATSE)", "Solters & Digney"

]

}

}

### countries

A list of ISO 3166 country codes for the countries in which the production companies for the title are based. For more details see [IMDb’s help site](https://help.imdb.com/article/contribution/titles/countries/GTSW4DN8H8LKCXER).

### episodeInfo

For titles that are episodes, this contains information about the series, such as the series title ID, sea- son number and episode number. It also includes the season and episode numbers where relevant.

#### Example

{

"episodeInfo":

{

"seriesTitleId": "tt0944947", "episodeNumber": 1,

"seasonNumber": 8

}

}

### seriesInfo

For titles that are series, this contains additional information about the series, such as the year it started airing, the year it finished airing (if it has finished), and a list of all the episode title IDs in the series ordered by episode number (e.g. season 1 episode 1, season 1 episode 2, etc.).

#### Example

{

"seriesInfo": {

"startYear": 2011,

"endYear": 2019, "episodeTitleIds": [

"tt1480055", "tt1668746", "tt1829962",

**...**

]

}

}

### episodeTitleIds

For titles which are series, the IMDb title IDs for all the episodes of that series.

### officialSiteLinks

A list of URLs (and optionally their link titles) linking to this title’s official website.

#### Example

{

"officialSiteLinks": [

{

"url": ["www.example.com/official/example-title"](http://www.example.com/official/example-title), "linkTitle": "Example official website for title"

},

**...**

]

}

### genres

A list of genres to which this title belongs. The full list of allowed genres and guidelines for how titles should be categorized can be found on [IMDb’s help site](https://help.imdb.com/article/contribution/titles/genres/GZDRMS6R742JRGAG). IMDb defines a limited list of genres but may add more in the future.

### image

A URL linking to an image associated with this title, such as a movie poster or still frame. Additionally includes the width and height of the image in pixels.

#### Example

{

"image": {

"url": "https://m.media-

↪ amazon.com/images/M/MV5BMWU4N2FjNzYtNTVkNC00NzQ0LTg0MjAtYTJlMjFhNGUxZDFmXkEyXkFq

"height": 1500,

"width": 974

}

}

### imdbUrl

A full URL to see the name or title on [www.imdb.com.](http://www.imdb.com/)

### isAdult

Whether or not this title contains adult content. Useful if you would like to filter out all adult content from your copy of the data set.

### keywords

A list of keywords associated with the title. More information about keywords and guidance for how they might be associated with at title can be found on [IMDb’s help site](https://help.imdb.com/article/contribution/titles/keywords/GXQ22G5Y72TH8MJ5).

### languages

A list of ISO 639 language codes for the languages spoken in this title, in order of frequency that they are spoken in the title. For more details see [IMDb’s help site](https://help.imdb.com/article/contribution/titles/languages/GSCRBZYEH5NX5SXL).

### locations

A list of locations where scenes from this title were filmed and optionally names or descriptions of the scenes which used that location.

#### Example

{

"locations": [

{

"scenes": ["studio"],

"place": "Fox Movietone Studio, New York, USA"

},

**...**

]

}

### movieConnections

A list of IMDb title IDs of other titles which have a connection to this title, and the type of connection, for example titles which reference or spoof this title. Optionally may include a description of the con- nection. A complete list of current connection types can be found on [IMDb’s help site](https://help.imdb.com/article/contribution/titles/movie-connections/GNUNL9W2FTZDGF4Y), although more may be added in future.

#### Example

{

"movieConnections": [

{

"type": "referenced\_in", "titleId": "tt2336547",

"text": "Jack criticizes the film for depicting 11 Americans being

↪

},

**...**

]

}

swayed by Jane Fonda's father"

### plot

A plot description of this title. Most plot descriptions will be just a couple of sentences long, however some may be longer, the ‘plot’ will contain the shortest of ‘plotShort’, ‘plotMedium’ or ‘plotLong’ and may be omitted. If you are displaying these plots you may need to consider truncation on longer plots.

### plotShort

A plot outline of this title, no longer than 239 characters. Plot outlines never contain spoilers.

### plotMedium

A plot summary of this title. Most plot summaries will be reasonably brief, a paragraph or two. If there are multiple plot summaries available on this title’s Plot page on [IMDb.com](https://www.imdb.com/), then the one provided here will have been selected to display prominently on the title’s main page by our users or manual vetting team.

### plotLong

A synopsis of this title. A long detailed description of the entire plot of the title.

### releaseDates

A list of the release dates (ISO 8601 date format) for this title, together with the region (an ISO 3166 country code) to which each release date applies.

Note that each release date may specify year, month and day (e.g. 1979-08-16), year and month (e.g. 1979-08) or only year (e.g. 1979).

#### Example

{

"releaseDates": [

{

"date": "1957-04",

"region": "GB"

},

{

"date": "1957-04-10",

"region": "US"

},

{

"date": "2016-02-24",

"region": "CZ"

},

**...**

]

}

### Production Status

A list of production statuses for this title in ascending order by date, with the last status being the current production status. The available statuses for in-production listings are available on [IMDb’s](https://help.imdb.com/article/contribution/titles/production-status/GBBBPNJHFZZKY7W9) [help site](https://help.imdb.com/article/contribution/titles/production-status/GBBBPNJHFZZKY7W9).

{

"productionStatus": [

**...**{

"updated": "2008-12-02", "status": "pre production"

},

{

"updated": "2009-10-24", "status": "filming"

},

{

"updated": "2011-04-17", "status": "released"

}

]

}

### runtimeMinutes

The running time of this title in minutes.

### taglines

A list of taglines for this title. A tagline is a short description or comment on a title that is often dis- played on posters. For additional details see [IMDb’s help site](https://help.imdb.com/article/contribution/titles/taglines/G6ZXC4NF9NUDSTZX).

### titleType

The type of this title, e.g. ‘movie’ or ‘episode’.

### imdbRating

The IMDb Rating for the title. The rating is between 1 and 10 and given to one decimal place. See [IMDb’s help site](https://help.imdb.com/article/imdb/track-movies-tv/weighted-average-ratings/GWT2DSBYVT2F25SK) for more information on how the rating is calculated. We also include the number of IMDb users who have voted on this title. A single IMDb user can cast a maximum of one vote. This field can be missing from an entry when we do not yet have an IMDb rating for the title in question. This can occur either because it does not yet have enough votes, or it has not yet been released. A TV series rating is not the weighted average of the ratings of individual episodes. Instead, customers vote separately for the rating of the series as a whole via each title’s series page.

#### Example

{

"rating": 8.9,

"numberOfVotes": 613399

}

### year

The year of the earliest release of this title globally.

## Data Dictionary - Box Office

### boxoffice\_title\_budgets\_v1

This file includes the reported production budget by title, where available.

#### budgetItemType

A description of the type of budget. The only valid value at this time is “production”.

#### amount

The budget for the associated titleId and budgetItemType, in USD.

### boxoffice\_title\_grosses\_v1

This file includes lifetime grosses for available titles, by area (see Appendix 1).

#### area

A code describing the area covered by this gross amount. See Appendix 1 for the full list of valid ar- eas.

#### grossToDate

The lifetime gross for the title within the relevant area, as most recently reported, in USD.

#### rank

The all-time rank for that title within the relevant area.

### boxoffice\_title\_opening\_weekends\_v1

This file includes opening weekend grosses for available titles, by area (see Appendix 1), along with the dates covered.

#### area

A code describing the area covered by this opening weekend. See Appendix 1 for the full list of valid areas.

#### startDate

The first day of the opening weekend for that title. Formatted as ISO 8601 (YYYY-MM-DD).

#### endDate

The last day of the opening weekend for that title. Formatted as ISO 8601 (YYYY-MM-DD).

#### ocassionId

The type of date span. The only valid value at this time is “weekend”.

#### gross

The gross for the title within the relevant area, between the start and end dates, in USD.

#### numTheaters

The theater count for the title within the relevant area, between the start and end dates.

### boxoffice\_title\_grosses\_timeseries\_v1

This file includes full time-series data by day, weekend, week, and more for available titles, by area (see Appendix 1), along with the dates covered and lifetime gross as of that time.

Note that not all titles will have complete coverage for every combination of area and occasionId. In other words for any particular area the value of grossToDate may not be equal to the sum of all gross values. They are best used in conjunction: gross gives you summary data while the time- series file allows you to drill down into more granular slices where available.

#### area

A code describing the area covered by this gross value. See Appendix 1 for the full list of valid areas.

#### startDate

The first day of the time span that this entry refers to. Formatted as ISO 8601 (YYYY-MM-DD).

#### endDate

The last day of the time span that this entry refers to. Formatted as ISO 8601 (YYYY-MM-DD).

#### ocassionId

The type of date span. Valid values include “Weekend”, “Weekly” and “Daily” as well as numerous special events and holidays.

#### gross

The gross for the title within the relevant area, between the start and end dates, in USD.

#### grossToDate

The lifetime gross for the title within the relevant area, as of the end date, in USD.

#### numTheaters

The theater count for the title within the relevant area, between the start and end dates.

#### rank

The rank for that title within the relevant area between the start and end dates.

## Querying Data in Amazon Athena Tables

### What is Amazon Athena?

Amazon Athena is an interactive query service that makes it easy to analyze data directly in Amazon Simple Storage Service (Amazon S3) using standard SQL. For more information, and for getting started with Athena, read the [user guide](https://docs.aws.amazon.com/athena/latest/ug/what-is.html).

### Getting Started

You first need to create a database in Athena. This process is documented in the [user guide](https://docs.aws.amazon.com/athena/latest/ug/what-is.html)

When you have a database, you’re ready to create a table that’s based on the dataset. You need to upload your dataset to Amazon S3. When you specify a location for your table you should use a trailing slash for your folder or bucket. Do not use file names or glob characters.

**Use:** s3://S3-BUCKET/S3-KEY/

**Don’t use:** s3://S3-BUCKET s3://S3-BUCKET/S3-KEY/\* s3://S3-BUCKET/S3-KEY/DATASET.JSONL.GZ

Athena will query all objects in the specified location so it is important that only one dataset is found at that path.

To create a table use the create table DDL statement found at the end of this document. Remember to set the location to the location of your dataset. If you do not need to query all of the columns in the table you can remove them from the create table DDL statement.

Now that you have a table created in Athena based on the data in Amazon S3, you can run queries on the table and see the results in Athena.

We have provided some example queries for common use cases.

### What Are the Highest-Rated Movies On IMDb?

IMDb user ratings can be found in the title essential dataset as part of the imdbRating structure.

select

tc.titleId, tc.originalTitle, tc.imdbRating.rating, tc.imdbRating.numberOfVotes

from

title\_essential\_v1 as tc where

tc.remappedTo is null and tc.titleType = 'movie' order by

tc.imdbRating.rating desc, tc.imdbRating.numberOfVotes desc

Running this query might return the following results:

|  |  |  |  |
| --- | --- | --- | --- |
| titleid | originaltitle | rating | numberofvotes |
| tt0050083 | 12 Angry Men | 8.9 | 616319 |
| tt0110413 | Léon: The Professional | 8.5 | 953632 |
| tt0064116 | Once Upon a Time in the West | 8.5 | 278746 |

### What Are the Title Texts for the Titles That a Person Is Known For?

The title IDs that a person is known for can be found in the name essential dataset as part of the knownFor array. To query this array it is necessary to flatten it into multiple rows using CROSS JOIN in conjunction with the UNNEST operator. To include the original title text it is necessary to JOIN the title essential dataset.

select

nc.nameId, nc.name, u\_knownFor.titleId, tc.originalTitle

from

name\_essential\_v1 as nc

cross join

unnest(nc.knownFor) with ordinality as t(u\_knownFor, ordinal)

join

title\_essential\_v1 as tc

on u\_knownFor.titleId = tc.titleId

where

nc.remappedTo is null order by

nc.nameId, ordinal

Running this query might return the following results:

|  |  |  |  |
| --- | --- | --- | --- |
| nameid | name | titleid | originaltitle |
| nm0000020 | Henry Fonda | tt0050083 | 12 Angry Men |
| nm0000020 | Henry Fonda | tt0082846 | On Golden Pond |
| nm0000020 | Henry Fonda | tt0032551 | The Grapes of Wrath |

### What Are the Names of the Principal Cast for a Title?

The name IDs for the principal cast for a title can be found in the title essential dataset as part of the principalCastMembers array. To query this array it is necessary to flatten it into multiple rows using CROSS JOIN in conjunction with the UNNEST operator. To include the name it is necessary to JOIN the name essential dataset.

select

tc.titleId, tc.originalTitle, nc.name

from

title\_essential\_v1 as tc

cross join

unnest(tc.principalCastMembers) with ordinality as t(u\_pcm, ordinal) join

name\_essential\_v1 as nc

on u\_pcm.nameId = nc.nameId

where

tc.remappedTo is null order by

tc.titleId, ordinal

Running this query on the sample dataset returns the following results:

|  |  |  |  |
| --- | --- | --- | --- |
| titleid | originaltitle | nameid | name |
| tt0050083 | 12 Angry Men | nm0000020 | Henry Fonda |
| tt0050083 | 12 Angry Men | nm0002011 | Lee J. Cobb |
| tt0050083 | 12 Angry Men | nm0000842 | Martin Balsam |

### What Awards was a Title Nominated for, and who were the Award Nominees?

select

tc.titleId, tc.originalTitle, tc\_awards.awardNominationId, tc\_awards.awardName, nc.nameId,

nc.name

from

name\_essential\_v1 as nc

cross join

unnest(nc.knownFor) with ordinality as t(u\_knownFor, ordinal)

join

title\_essential\_v1 as tc

on u\_knownFor.titleId = tc.titleId

cross join

unnest(tc.awards) as t(tc\_awards) cross join

unnest(nc.awards) as t(nc\_awards) where

tc\_awards.awardNominationid = nc\_awards.awardNominationId

Running this query might return the following results:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| titleId | originalTitle | awardNominationId | awardName | nameId | name |
| tt0032551 | The Grapes of Wrath | an0052939 | Oscar | nm0002034 | Jane Darwell |
| tt0032551 | The Grapes of Wrath | an0829169 | NBR award | nm0002034 | Jane Darwell |
| tt0032551 | The Grapes of Wrath | an0052926 | Oscar | nm0000020 | Henry Fonda |

### What Are the Title Texts for Episodes of a Series?

The title IDs for episodes that are part of a series can be found in the title essential dataset as part of the episodeTitleIds array. To query this array it is necessary to flatten it into multiple rows using CROSS JOIN in conjunction with the UNNEST operator. To include the title text it is necessary to JOIN the title essential dataset.

select

tc\_series.titleId, tc\_series.originalTitle, tc\_episode.titleId, tc\_episode.originalTitle, tc\_episode.episodeInfo.seasonNumber, tc\_episode.episodeInfo.episodeNumber

from

title\_essential\_v1 as tc\_series

cross join

unnest(tc\_series.seriesInfo.episodeTitleIds) as t(u\_eti)

join

title\_essential\_v1 as tc\_episode on u\_eti = tc\_episode.titleId

where

tc\_series.remappedTo is null order by

tc\_series.titleId, tc\_episode.episodeInfo.seasonNumber, tc\_episode.episodeInfo.episodeNumber

Running this query might return the following results:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| titleid | originaltitle | titleid | originaltitle | seasonnumber | episodenumber |
| tt5491994 | Planet Earth II | tt6142646 | Islands | 1 | 1 |
| tt5491994 | Planet Earth II | tt6209126 | Mountains | 1 | 2 |
| tt5491994 | Planet Earth II | tt6209130 | Jungles | 1 | 3 |

### What Is the Title Text of the Series for an Episode Title?

The title ID for a series that an episode is part of can be found in the title essential dataset as part of the

episodeInfo structure. To include the title text it is necessary to JOIN the title essential dataset.

select

tc\_episode.titleId, tc\_episode.originalTitle, tc\_series.titleId, tc\_series.originalTitle, tc\_episode.episodeInfo.seasonNumber, tc\_episode.episodeInfo.episodeNumber

from join

title\_essential\_v1 as tc\_episode

title\_essential\_v1 as tc\_series

on tc\_episode.episodeInfo.seriesTitleId = tc\_series.titleId

where

tc\_episode.remappedTo is null order by

tc\_episode.titleId

Running this query might return the following results:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| titleid | originaltitle | titleid | originaltitle | seasonnumber | episodenumber |
| tt6142646 | Islands | tt5491994 | Planet Earth II | 1 | 1 |
| tt6209126 | Mountains | tt5491994 | Planet Earth II | 1 | 2 |
| tt6209130 | Jungles | tt5491994 | Planet Earth II | 1 | 3 |

## Appendix 1: Box Office Mojo Areas

### Area Rollups and Special Areas

Area

abbreviation Full area description Notes

XDOM Domestic Reporting includes

US and Canada

XWW Worldwide

XNDOM International

XA2 Southern, Eastern, and Western Africa

XA3 South Africa and Nigeria

XB2 Baltic States

XBA Bosnia

XC3 Czech Republic and Slovakia

XC4 Belize, Costa Rica, El Salvador, Guatemala, Honduras, Nicaragua, Panama and Paraguay

XCN Aruba, Belize, Bolivia, Bonaire, Costa Rica, Curacao, and Dominican Republic

XEW Ghana, Kenya, Nigeria, Tanzania, Zambia and Uganda XG2 West Germany

XGC Greece and Cyprus

XKN Kenya, Somalia, Tanzania and Uganda XL2 Latin America

XL3 Lebanon and United Arab Emirates

XM2 Middle East

XM7 Middle East including United Arab Emirates XMO Miscellaneous Middle East

XR2 Russia and Commonwealth of Independent States XRE Reunion

Area

abbreviation Full area description Notes

XRK Russia, Commonwealth of Independent States and Ukraine

XS1 Switzerland (German)

XS2 Serbia and Montenegro

XS3 Switzerland (French)

XS4 Switzerland (Italian)

XS5 Switzerland (French/Italian)

XT2 Antigua and Barbuda, Barbados, Dominica, Grenada, Guyana, St. Lucia, and St. Vincent

XW2 West Indies

### Individual Areas

Area abbreviation Full area description Notes AD Andorra

AD Andorra

AE United Arab Emirates

AF Afghanistan

AG Antigua

AL Albania

AM Armenia

AN Netherlands Antilles

AO Angola Included in PT

AQ Antarctica

AR Argentina

AT Austria

AU Australia

AW Aruba

|  |  |  |
| --- | --- | --- |
| Area abbreviation | Full area description | Notes |
| AZ | Azerbaijan |  |
| BT | Bhutan |  |
| BA | Bosnia and Herzegovina |  |
| BB | Barbados |  |
| BE | Belgium | Reporting includes Luxembourg |
| BH | Bahrain |  |
| BM | Bermuda |  |
| BD | Bangladesh |  |
| BO | Bolivia |  |
| BR | Brazil |  |
| BS | Bahamas |  |
| BW | Botswana |  |
| BY | Belarus |  |
| BZ | Belize |  |
| CH | Switzerland |  |
| CO | Colombia |  |
| CU | Cuba |  |
| CA | Canada |  |
| CN | China |  |
| CL | Chile |  |
| CR | Costa Rica |  |
| CW | Curacao |  |
| CY | Cyprus |  |
| CZ | Czech Republic |  |
| DE | Germany |  |
| DJ | Djibouti |  |
| DK | Denmark |  |

|  |  |  |
| --- | --- | --- |
| Area abbreviation | Full area description | Notes |
| DO | Dominican Republic |  |
| DZ | Algeria | Included in FR |
| KE | Kenya |  |
| EC | Ecuador |  |
| EE | Estonia |  |
| EG | Egypt |  |
| ES | Spain |  |
| ET | Ethopia |  |
| FI | Finland |  |
| FR | France | Reporting includes Algeria, Monaco, Morocco, and Tunisia |
| GH | Ghana |  |
| GR | Greece |  |
| GT | Guatemala |  |
| GU | Guinea |  |
| NL | Netherlands |  |
| HK | Hong Kong |  |
| HN | Honduras |  |
| HR | Croatia |  |
| HU | Hungary |  |
| ID | Indonesia |  |
| IE | Ireland | Included in GB |
| IL | Israel |  |
| IN | India |  |
| IQ | Iraq |  |
| IR | Iran |  |
| IS | Iceland |  |
| IT | Italy |  |

|  |  |  |
| --- | --- | --- |
| Area abbreviation | Full area description | Notes |
| JM | Jamaica |  |
| JO | Jordan |  |
| JP | Japan |  |
| KH | Cambodia |  |
| KR | South Korea |  |
| KW | Kuwait |  |
| KZ | Kazakhstan |  |
| LI | Liechtenstein |  |
| LA | Laos |  |
| LB | Lebanon |  |
| LT | Lithuania |  |
| LV | Latvia |  |
| LU | Luxembourg | Included in BE |
| MA | Morocco | Included in FR |
| MK | North Macedonia |  |
| MT | Malta | Included in GB |
| MN | Mongolia |  |
| MU | Mauritius |  |
| MX | Mexico |  |
| MY | Malaysia |  |
| MZ | Mozambique |  |
| NC | New Caledonia |  |
| NG | Nigeria |  |
| NI | Nicaragua |  |
| NO | Norway |  |
| NZ | New Zealand | Reporting includes Fiji |
| OM | Oman |  |

|  |  |  |
| --- | --- | --- |
| Area abbreviation | Full area description | Notes |
| PA | Panama |  |
| PS | Palestine |  |
| PE | Peru |  |
| PY | Paraguay |  |
| PH | Philippines |  |
| PK | Pakistan |  |
| PL | Poland |  |
| PR | Puerto Rico |  |
| PT | Portugal | Reporting includes Angola |
| QA | Qatar |  |
| RO | Romania |  |
| RU | Russia |  |
| RW | Rwanda |  |
| SU | HH Soviet Union |  |
| SN | Senegal |  |
| SA | Saudi Arabia |  |
| SE | Sweden |  |
| SG | Singapore |  |
| SI | Slovenia |  |
| SV | El Salvador |  |
| LK | Sri Lanka |  |
| SK | Slovakia |  |
| SR | Suriname |  |
| SY | Syria |  |
| TH | Thailand |  |
| TR | Turkey |  |
| TT | Trinidad and Tobago |  |

|  |  |  |
| --- | --- | --- |
| Area abbreviation | Full area description | Notes |
| TN | Tunisia | Included in FR |
| TW | Taiwan |  |
| UT | Uruguay |  |
| GB | Great Britain | Reporting includes, UK, Ireland, and Malta |
| UZ | Uzbekistan |  |
| VE | Venezuela |  |
| YU | Yugoslavia |  |
| ZA | South Africa |  |