# NUTFLUX™

## The Home of the Knowledgeable Movie Nut

**Project Whitepaper**

*Chadwick P. Knutsack, Jnr.*

*Chairman Emeritus,*




CONFIDENTIAL

## 1. A New Era In Streaming Content

In recent years, NUTFLUX™ has become the dominant streaming platform for TV and film content, with its market dominance only strengthened by the COVID-19 pandemic. However, this dominance has drawn new attention from large content providers eager to host their own content online. Although NUTFLUX is a serious producer of original content in its own right, the company must seek to improve its offering and differentiate its brand not only with sustained output of movie and TV products, but with additional features that enrich the product experience, so as to attract new subscribers and to strengthen the loyalties of existing users.

In the coming months, the company will pursue multiple upgrade paths with complementary offerings, but this current document focuses on just one: a data-based, knowledge-rich offering that will make the NUTFLUX platform far more attractive to cineastes (that is, "movie nuts", our core constituency) and to those whose interests are less academic and more geared toward pop-culture "gossip." To source the copious data and rich knowledge-structures to make this offering a reality, NUTFLUX has proposed a partnership to the *Internet Movie Database*, or IMDb, but that company feels that the purity of its mission may be tarnished, or otherwise diluted, by the addition of tabloid-style information about movie stars and their personal relationships to their existing databases. As a consequence, NUTFLUX has decided to go it alone, by replicating the key data offerings of the IMDb and by integrating the additional layers of new content itself. The new DB will augment the NUTFLUX platform in two guises: the standard user experience will be enriched by the addition of IMDb-style factoids in the content browser, such as film ratings (scores out of 10 based on the aggregation of viewer ratings), cast information (main stars, guest directors, etc.), and relevant movie trivia; and the "power user" experience, called NUTFLUX-*Pro*, in which key tables and views are exposed to query access in SQL, so that professional users can obtain specific answers to complex queries (such as "What was Matt Damon's first credited role in a comedy directed by a woman?").

What relational form will these additions take? As RDBM foundations, they do not constitute a total solution or full application. Rather, they provide tables and

views for organizing NUTFLUX data in a clean, consistent and normalized fashion that permits easy and efficient access (via queries) and safe modification (via update & insert statements). You will also define sample queries that show your database design in action, and also define a variety of procedural elements to achieve ends that are not easily or conveniently realized in pure SQL terms. Your offering will be evaluated using a rubric that is provided in a separate document. Be sure to read that document to inform yourself of what is actually needed here.

## 2. Replicating the IMDb Offering

Since the IMDb is unwilling to work with NUTFLUX on its new knowledge-based features, we shall have to replicate for ourselves the key data elements in which a data-rich offering will be anchored. The most basic and necessary information concerns the creative talent behind each film or TV show that a user may wish to watch on our platform: the stars, the directors, the writers, the guest stars and guest directors (for TV shows), a plot synopsis for each, and perhaps nuggets of trivia, or some well-known quotes ("these are not the droids you are looking for", or "You're going to need a bigger boat", or "You were only supposed to Blow THE BLOODY DOOR OFF!"), to stimulate interest, stir feelings of nostalgia, and make a piece of content stickier and more enticing to casual browsers.

To appreciate what tables and views are required for the core database, one must first make a reasonable prediction of what queries will be issued against it, both by the standard browsing platform ("standard users" issue no queries for themselves, but are served data in the context of their browsing activities) and by the professional "power users" who will access non-private areas of the DB directly using their own SQL formulations. Let's consider the *standard user* first. As such users browse the NUTFLUX content catalogue, they will be presented with important information which may sway their decisions as to what to watch. What draws a standard user to a given piece of content? Most obviously, it is the *stars* of the show, the actors or presenters who carry the production. In some cases a streaming platform will offer a short list of actors in alphabetical order, prioritizing those who names begin with A and B over those that actually attract viewers! Ideally, the main actors will be listed first, as these reflect the greatest

investment by the studio and attract the most attention from potential viewers. More sophisticated standard users will base their viewing decisions on who has directed the production. Those who are most sophisticated still may decide on the basis of the *writer* (or *writers*). Since many directors and writers of quality are not household names, it may help if the platform includes a note of the form "from the director of <big hit>" or "from the same writer as <classic film>". If a user has given a good rating (a "thumbs up") to another piece of content with the same star, director, producer or writer, this may also be a useful note to include. What other pieces of actionable intelligence might the browsing platform use to attract a listless browser's interest? Perhaps one of the stars of the piece has won an Oscar, or has been nominated for this very role? The adman and the huckster are one when it comes to pushing content in a crowded marketplace of choices.

Now, what of the needs of the professional *power user*? These are not passive consumers of data, but active seekers of factual answers to targeted questions. Such users decide on the questions to ask, the information to be returned, and the format of that information. In other words, they use all the affordances that are available to them via SQL and the tables/views which the platform exposes. They might seek specific content that addresses specific needs, with queries like "show me all films starring either Ben Affleck or Matt Damon but not both." They may seek to resolve certain factual questions, such as the date of a specific film, or the earliest date at which a certain fact became true (such as the year in which a woman first won the academy award for best director). They will satisfy these needs using complex table joins that integrate diverse information sources and views onto the underlying data, and they may create views of their very own for future use (this can be marketed as a premium service for premium users). Date information will be of particular information to these users, in ways that do not appeal to the casual or standard user. Information regarding awards (winners and nominees) will be of some interest to both standard and power users, but only the latter will be particularly interested in the years of those awards. For example, a power user may ask "which actors who have played *The Joker* have won an academy award?" or, for a more targeted answer, "which actors who have played *The Joker* have *not* won an academy award?". In these queries, a user

will want more than the name of the actor; they will want the name of the film in which they have played *The Joker*, the name of the film for which they won the award, and the year of release for each of them. Notice that these queries also make reference to the *names* of the characters played by actors in a production.

A cross-cutting concern to both standard and power users is ambiguity: many films have non-unique names, either by chance or by design. Some, for instance, are remakes or reboots of others, while others draw their inspiration, and titles, from the same source material. For instance, *Dune* (1984), as directed by David Lynch, is based on the same book (of the same name, from author Frank Herbert) as *Dune* (2021), as directed by Denis Villeneauve. It seems clear, then, that titles are not unique keys for referencing content in the database. We do not expect the same ambiguity in the naming of actors and directors, since professional unions insist that their members adopt unique names (e.g., the actor Micheal Keaton is actually named Michael Douglas; he changed it for obvious reasons). Sometimes we see the results of this ambiguity, as when a streaming service or TV service shows the wrong image, poster, plot summary or cast for a scheduled movie (see, for example, Virgin Media's frequent mistakes, or those of Amazon Prime). For the power user, it may be important to not only distinguish between different forms of the same material (e.g., *Dune* in 1984 and *Dune* in 2021) but to capture their intrinsic relationship. To a power user, *Dune* (2021) is *not* a remake of *Dune* (1984). Rather, the latter is another interpretation of the same source material. Clearly, our scope for putting *NUTFLUX*'s platform on a knowledge-based footing is immense, not just for power users, but for ordinary users who must be given pertinent information at the right time, or for whom a useful recommendation is grounded in a proper understanding of what is actually being recommended. The success of this data-rich augmentation relies not just on the scale of the data that is added, but on the useful combinatorial possibilities you can demonstrate for it.

## 3. Behind The Candelabra

So far we have considered the kinds of information that one would expect to find on IMDb.com, both from the perspective of casual users (who are *pushed* the data as needed) and power users (who *pull* the data they desire). However, NUTFLUX

aims to augment this domain content with social content, of the kind that one would find not in specialist magazines but supermarket tabloids like the *National Enquirer*. This content might be considered gossip and scuttlebutt, but it adds a social dimension that attracts a new level of interest from users, or a whole new demographic of users. Many of our users, or potential subscribers, are genuinely interested in the romantic entanglements of the stars who make their favorite content, such as who is married to whom, who is dating whom, who is divorced from whom, and so on. Fundamentally, our users are interested in connections: in seeing connections and in making them too. Users might search, for instance, for films where there is a real-world romantic relationship between the leading actors, paying special attention to the date of the film and the start date of the relationship (e.g., did two actors marry after appearing together, or even divorce after doing so, or perhaps divorce after appearing with another co-star that they then became involved with?). The social dimension is interesting in its own right, but the corresponding tables are most useful when combined with the content database to pose queries than combine fact and fiction, or business and romance.

Making a film together is itself a kind of social entanglement. Just think of the game Six Degrees of Kevin Bacon, in which actors are assigned a *Bacon Number* that corresponds to their professional distance from this prolific actor. Any actor who appears in the same film as Bacon earns a Bacon number of 1, while those who appear in a film with someone who has a Bacon number of *N* will receive a Bacon number of *N+1* (unless they have another, shorter path to the actor). This is just the kind of intellectual exercise that movie nuts can use the NUTFLIX DB to play for themselves, by posing the appropriate queries in SQL.

The notion of a "role" is also important when querying a database of film data. Perhaps you are interested in finding actors who became directors, or vice versa, or in finding actors who have directed themselves, or in directors who take up cameo roles in each other's films? Perhaps you are looking for pairs of actors who work together a great deal, or in finding romantically entangled stars who never work together. Perhaps you are looking for actors who made the transition from TV actor to movie actor, or vice versa, for actors who built their careers in cinema but who now appear primarily on the small screen. If the database stores

not just the names of the characters played by actors, but the general category of those characters (e.g., hero, villain, antihero, love interest, etc.) or a more specific category still (e.g., detective, mathematician, reporter, etc.) then rich queries that take advantage of these categorizations can be posed by our power users. Which actor has played both a detective and a mathematician? Which Oscar-winning actors have played superheroes (or starred in superhero films)? Which actor has won an Oscar for playing the villain? Are actors more likely to win Oscars for playing the villain than for playing the hero? Or for playing real-life people from history versus entirely fictional creations? Has anyone ever won an Oscar for just providing the voice of a character? How many characters have been portrayed on screen in multiple outings by more than one Oscar winner? These are the kinds of cross-cutting queries that curious power users love to pose, and just the kind of queries that our newly data-enriched NUTFLUX platform will support.

## 4. Creative Content Classification

The kinds of technical role in the production of content are jealously guarded and well-defined by the professional unions. There is little ambiguity about what constitutes a director, an actor or a writer, though when it comes to awards, one can quibble about whether a part is a leading role or a supporting one. Anyone who has ever used the NUTFLUX platform, or that of our competitors (e.g., of *Netflix*, *Amazon Prime*, *Disney*+) will know that content is typically organized by category, and these categories are highly conventionalized. The most timeworn categories cover the *Suspense*, *Action*, *Horror*, *Comedy, Romance, Fantasy* and *Sci-Fi* genres, but there are broad crossover categories too, such as *RomCom* (a film that is both *Comedy* and *Romance*), futuristic thriller (*Sci-Fi* and *Action*), and even *Comedy Horror* (e.g., think of the *Hellboy* or *Scream* franchises). Indeed, the most creative productions – and thus the ones most likely to appeal to hardcore fans – straddle the boundaries of conventional classifications. For instance, is 1988's *Die Hard* a Christmas movie? It has all the trappings of a festive film, from its plot (a daring heist on Christmas Eve) to its musical choices, to its Scrooge-like arc from humbug to blessings, to a love interest named *Holly*. In much the same way, 1971's *Willy Wonka & the Chocolate Factory* is very clearly an Easter film.

The same creative ambiguity can also be found in how we categorize the roles played by an actor. Think of the genre-specific role of "Bond girl," made famous by Ursula Andress in 1962's *Dr. No*, by Honor Blackman in 1964's *Goldfinger*, and by Diana Rigg in 1969's *On Her Majesty's Secret Service*. What defines a Bond girl? Is it enough to be female and central to the plot? Is Dame Judy Dench a Bond girl? When it comes to creative categorization, or to any situation where we have to be flexible in how we use and apply categories, the knowledge-based approach truly comes into its own. Ultimately, knowledge has as much to do with what we assume as with what we actually "know," and the knowledge-based approach to information architecture is defined by an explicit articulation of what is assumed and of what is accepted as fact. Our power users will want to be creative in how they use the knowledge of the NUTFLUX platform in combination with their own assumptions, as embodied in their queries, views, and so on. So this is an area in which you, as an information architect and a knowledge engineer, can do most to impress with your category-defying insights and playful cross-cutting designs.

## 5. Project Reporting

An accompanying document provides a template for the report you will submit alongside your SQL definitions for this database. In addition, a marking guide is provided so as to help you focus your efforts on what is expected of you and in what areas you can expect to receive the most scrutiny and the most marks.

Do not skimp on the report, or on the sections that you believe are mere filler. Your report is a proposal – a whitepaper of your own – that will go to the board of NUTFLUX Inc. for a thorough evaluation. As a DBA it is your responsibility to not merely design a database, but to argue for your design, making it clear why certain design choices were taken. A DBA's job requires *hard* skills (for SQL and related coding tasks) and *soft* skills (to explain and justify the former, to your team, to clients, and to non-DB staff).

Your report should contain your thoughts and your insights. It should not just recycle the content of this preliminary whitepaper, or just fill space with generic discussions of the role of a database in the enterprise. Be specific and on-point, and outline your vision for the project and your database clearly. Be sure to read

this white paper carefully, so as to understand the demands that will be placed upon your design. You should also be careful to communicate this design well, in the report that will accompany your SQL content. Motivate each of your design decisions, explain why they are sound (e.g., by demonstrating the normalization of all tables up to BCNF), and provide ample examples of the design being used to support queries and procedures that are of obvious use in the NUTFLUX pipeline.

## 6. Conclusions

When I was made Chairman Emeritus of NUTFLUX Inc. in 2019, I naturally asked my successor as chair of the board, "What does 'Emeritus' mean?" I was given one of the frankest answers of my career, the candour of which still rankles and inspires: "*Well, Chad*", I was told, "The *'E' means 'exit'. So you're out! And the 'meritus' means you bloody well deserve it!*" I fear I had spent far too much of the company's precious reserves on expensive fan-service vanity projects – does anyone now remember *Predator Vs. Tellytubbies*? – and far too little on growing our platform's core functionality. But this project gives you the opportunity to rectify my biggest mistake, by creating an information architecture to carry the company and its content offerings into a more competitive and data-rich future.

So what is expected of you, on a project my erstwhile colleagues of the board have nicknamed "*The Knutjob*"? At this stage, no one expects a complete solution. Rather, you are to view this project as an opportunity to paint your vision of the information architecture of the enhanced NUTFLUX platform. What is required is ample evidence of this vision, not a done deal, so tables need only be fleshed out to work as illustrations. Ensure each table has at least 10 entries to showcase its purpose, that each view has enough to showcase its value, that any procedural elements (triggers, procedures, or functions) perform a useful service well, and that your sample queries are neither overly simple or pointlessly complex. For these queries, you may take inspiration from the suggestions discussed here, but do also use this opportunity to show off your own imagination. As to the number of demonstrations in each case, of queries, views, procedural items and whatnot, I myself have always been a believer in the "rule of four." So, as I take my leave of you now, I wish you the very best on your NUTFLUX adventure. *Da-Dummm*!